Spatial smoothing in statistical regression models

(Alternative Format Thesis)

Dupont, Emiko

Award date:
2021

Awarding institution:
University of Bath

Spatial smoothing in statistical regression models

submitted by

Emiko Dupont

for the degree of *Doctor of Philosophy*

of the

University of Bath

Department of Mathematical Sciences

April 2021

COPYRIGHT

Declaration of any previous Submission of the Work

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

.....

Emiko Dupont

Declaration of Authorship

I am the author of this thesis, and the work described therein was carried out by myself personally, with the exception of Papers 1 and 2, where part of the work was carried out by other researchers as detailed in the statements of authorship.

.....

Emiko Dupont

Summary

Spatial regression models are commonly used in applied statistics to model data collected at different spatial locations. Such models use spatial random effects to account for residual spatial correlation in the response variable and result in fitted values that are, to some degree, smoothed across the spatial domain of the data. The main focus of this thesis is a problem known as spatial confounding, which causes covariate effect estimates in spatial models to be unreliable. By investigating the estimation theory in a commonly used spatial model formulation based on thin plate splines, we gain a deeper understanding of the problem and the existing methodology. Using this, we develop a novel and easily implementable method for avoiding spatial confounding in practice. Moreover, we include some initial analysis on spatial confounding in models with non-linear covariate effects; an area that has not yet been explored in the literature. The thesis also contains another project within the field of spatial statistics. Here, spatial modelling techniques are used to develop a method for detecting spatially coherent trends in environmental time series data. Specifically, we model river flow data from gauging stations across Great Britain. Using our methodology, we are able to verify, for the first time, a significant upward trend in flood risk over time and identify the regions with the largest trends.

Acknowledgements

Firstly, I would like to thank my examiners, Julian and Janine, for helpful comments and encouragement, and for making it possible to finalise my thesis under the cloud of yet another illness.

Much of my time in SAMBa was dominated by Alastair's crippling and poorly understood illness. Due to the treatment, generosity and ingenuity of Orlando, Margarida, Sue, Trevor, Karen, Ana and Marina, this is thankfully behind us, and I am forever grateful to them for that. I wholeheartedly thank SAMBa, particularly, Paul and Susie, for the support throughout this period and my supervisors Nicole, Simon, Ilaria and Matt for all your insights and guidance and for sticking with me through uncertain times. Thank you also to Rob, Kei, Karim, Hadid, Catherine, Kirsten and Kim for your friendship and for providing support and childcare that made seemingly impossible logistics possible.

I feel privileged to have been part of a wonderful and inclusive student and research community whose energy, encouragement and coffee breaks have meant the world to me. In particular, Aoibheann, Ben, Hayley, Connor, Matt, Elizabeth, Marcus, Tsogii, Will, Level 1 coffee, Level 2 coffee, Level 5 coffee, Strike office, COffice-19, Sanne and Thomas - thank you all for making my working environment a happy place to be. Special thanks go to Jack, for your help with everything IT and for being there when things were really tough; Kate, for your friendship throughout; and James, for being a great listener and compassionate friend. And Malena, I am truly grateful for everything you've done, not just for me, but for my family as well; I honestly, don't know what I would have done without you!

Finally, my deepest gratitude goes to my amazing family. To my sister, Yoko, for always being there for me. To my mum, who more than anyone understands what we went through, thank you for all your help and support. To my dad, a source of inspiration and wisdom, even now that he is gone. To Alastair for your strength, perseverance and optimism that got us through the most challenging of times; thank you for being who you are. And to Naomi and Logan, my little rays of sunshine; you make everything worthwhile and, no matter what happens, you always make me smile.

Contents

Chapter 1

Introduction

Spatial regression models are used in applied statistics to model data collected at different spatial locations. Such models use spatial random effects to account for residual spatial correlation in the response variable which arises, for example, as a result of unmeasured or unknown spatially dependent covariates that have not been included in the model. Spatial random effects reflect the expectation that observations close to each other are more likely to be similar than those far apart and result in fitted values that are, to some degree, smoothed across the spatial domain of the data. Spatial statistics has seen a rapid development over the last few decades, and technological and computational advances in the field have meant that relatively sophisticated models can now be fitted to data sets using standard software packages. This has made spatial modelling an increasingly common tool. This thesis explores some of the underlying methodology and includes two papers (one under review and one published) which we refer to throughout as Paper 1 and Paper 2.

Spatial regression is often used for prediction, i.e. estimating the values of the response variable in locations where we have no observations. But like other regression models, spatial models are also used for assessing the effect of individual covariates on the response variable. The bulk of the work in this thesis, including Paper 1, focusses on a problem known as spatial confounding which arises in this latter context. Random effects in regression models are usually assumed to be independent from the covariates in the model. But spatial random effects typically have elements of collinearity with spatially dependent covariates, and therefore they can interfere with the effect estimates of interest, making these estimates unreliable. The issue of spatial confounding was first identified by Clayton et al. [1993] and is analysed further in Reich et al. [2006], Hodges and Reich [2010], Paciorek [2010], Hanks et al. [2015], Page et al. [2017]. A well-known example in Reich et al. [2006] illustrates the problem: in a Poisson regression model for assessing the effect of socio-economic status on stomach cancer incidence in the municipalities of Slovenia, an initial regression without spatial random effects shows that the covariate effect is negative and significant. But when spatial random effects are added to the model, the covariate effect becomes close to zero and is no longer significant. This behaviour makes statistical inference difficult.

As spatial models are usually complex, explicit derivations of the effect estimates are not straightforward and, therefore, the underlying mechanism for spatial confounding issues is not fully understood. The most commonly used methods for dealing with the problem are based on orthogonalisation [Reich et al., 2006, Hanks et al., 2015, Hughes and Haran, 2013, Pereira et al., 2020, Adin et al., 2020], i.e. restricting the spatial random effects to the orthogonal complement of the covariates. This approach, known as restricted spatial regression (RSR), directly eliminates collinearity in the model matrix and results in covariate effect estimates that agree with the null model, i.e. the model with no spatial effects. But as others have pointed out, RSR can actually lead to significant bias in the effect estimates unless the unmeasured spatial effects are independent of the covariates in question; usually an unrealistic assumption [Khan and Calder, 2020, Sørbye et al., 2019]. This is because the RSR estimate reflects, not only the effect of the covariate, but also any unmeasured spatial effects that are associated with the covariate. Another proposed method is the geoadditive structural

equations model (gSEM) [Thaden and Kneib, 2018]. Here, spatial dependence is regressed away from both the response and the covariates, and a regression involving the residuals only is used to identify the original covariate effects. Simulations show that, using this approach, the bias in the covariate effect estimates of the spatial model is broadly removed, however, it is not immediately clear why the method works and it seems undesirable to remove all spatial information from the modelling.

The main objective for this PhD has been to gain a better theoretical understanding of spatial confounding; why it happens and what can be done to avoid it.

One notable observation, which is not usually highlighted in the literature, is that spatial confounding issues depend on the structure of the covariate of interest. A common assumption is that the covariate, like the response variable, has a spatially determined covariance structure. Intuitively, this means that the spatial model cannot distinguish the covariate from an unmeasured spatial effect due to the collinearity in the model matrix. The apportionment of effects between the covariate and spatial parts of the model may therefore be somewhat arbitrary and lead to bias in the covariate effect estimate. Paciorek [2010] and Page et al. [2017] studied the effect estimates in the spatial model under this scenario and show that the size of the bias depends on the relative spatial scales of the covariate and spatial effects and, when the spatial scales agree, the bias is the same as that of RSR (and, hence, the null model). Thus, while the estimate in the spatial model differs from RSR, it may be just as biased.

However, in many practical applications covariates are spatially dependent but not fully determined by spatial location. This distinction may seem subtle, but is important, as non-spatial information in the covariate can be used to distinguish it from the spatial effects without the need for considering differences in spatial scales. Our analysis shows that, in this case, the spatial model may still have significantly biased covariate effect estimates, however, the bias is a result of the combination of collinearity and spatial smoothing, rather than collinearity alone. Moreover, we show that such bias can be avoided in a relatively straightforward way. Adopting a thin plate spline formulation of the spatial model, we are able to write down explicit expressions for the effect estimates and analyse their behaviour using linear algebra. This formulation of the model is also easily implemented in the `R`-package `mgcv`. Our analysis lead us to discover a crucial link between spatial confounding in this context and the theoretical work of Rice [1986] and Chen and Shiau [1991] who studied the behaviour of effect estimates in semiparametric models where the domain of the spline (in our case, the spatial domain) is one-dimensional. A generalisation of this work to arbitrary spatial dimensions lead to the results in Paper 1, which contains some of the main results of this PhD.

Paper 1 provides a novel theoretically-backed method, spatial+, for avoiding spatial confounding bias when a covariate is not fully determined by spatial location. Through asymptotic analysis (as the number of fitted data points $n \to \infty$) of model estimates in the thin plate spline formulation, we show that smoothing of the spatial effect in the spatial model leads to disproportionate bias in covariate effect estimates, while the spatial+ model is able to capture the true effects with negligable bias. A simulation study illustrates that the method works and how it compares to existing methods. We also apply spatial+ to a data example using forestry data to assess the effect of temperature on tree health. While the main derivations in the paper are for the case where the response distribution is Gaussian, we show that the method extends to any response distribution from the exponential family of distributions.

The models considered in Paper 1 assume that the effect of the covariate of interest on the response variable is linear. However, another advantage of the thin plate spline formulation of the spatial model is that it fits into the framework of generalized additive models (GAMs) (see Hastie and Tibshirani [1990], Wood [2017]). This allows us to consider spatial models with possibly non-linear covariate effects. More specifically, the dependence of the response variable on the covariates are represented by unknown smooth functions that are estimated from the data. Spatial confounding in this context does not appear to have been studied in any detail in the literature. Our initial investigations show that the problem is still clearly present for these models. However, the estimation theory is more complex and our proposed solution, spatial+, which assumes linear covariate effects, does not directly generalise to this case. This would, however, be an interesting direction of future work.

Finally, Paper 2 [Prosdocimi et al., 2019] was another project for the PhD, within the area of spatial statistics, but unrelated to spatial confounding. In this project, spatial modelling is used for trend detection in time series data. Specifically, we model river flow data from gauging stations across Great Britain to investigate whether there is an upward trend in the annual maximum flow data series, i.e. whether flood risk is increasing over time. Here the spatial model is used for the purpose of data pooling. The data series at each location is relatively short, making it difficult to detect a significant statistical signal. Indeed, while the frequency of major flood events in recent years as well as climate models suggest there is likely to be an upward time trend, this could not previously be verified by the data. However, by modelling the data from all gauging stations in a single spatial model, information can be shared between different locations. Using this, the statistical signal is enhanced and we are able to detect, for the first time, a significant upward trend in flood risk over time. Moreover, the spatial model identifies the regions with the strongest trends. While the emphasis of the paper is the particular application to river flows, the method could easily be used for other applications as well.

The rest of this thesis is organised as follows. Chapter 2 summarises our work on spatial confounding which formed the basis for Paper 1 and the spatial+ approach. We have also included an initial investigation into spatial confounding in models with non-linear covariate effects. Chapters 3 and 4 introduce and discuss Papers 1 and 2, respectively. Finally, Chapter 5 sets out the overall conclusions of the thesis and a number of directions for future research.

Chapter 2

Spatial confounding

In this chapter we summarise our work on spatial confounding that formed the foundation for the results in Paper 1. We introduce the thin plate spline formulation of the spatial model which allows us to analyse the behaviour of effect estimates using linear algebra. A detailed asymptotic analysis of these estimates, as well as results of simulations and an application to a data example, are included in Paper 1. However, in this chapter, the focus is on understanding the underlying mechanisms behind the confounding problems, namely, collinearity and spatial smoothing. As in Paper 1, most of our analysis assumes a Gaussian response distribution, however, we also show how the results generalise to response distributions from the exponential family of distributions. As part of our investigation, we implement and analyse the method of RSR. We have also included some initial investigations into spatial confounding in spatial models with non-linear covariate effects.

Spatial modelling using thin plate splines

Suppose we have response data $\mathbf{y} = (y_1, \ldots, y_n)^T$ and covariate data $\mathbf{x} = (x_1, \ldots, x_n)^T$ measured at spatial locations $\mathbf{t}_1, \ldots, \mathbf{t}_n$ in $\mathbb{R}^d$ with dimension $d \geq 1$. Our starting point is the null model

$$y_i = \beta x_i + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2) \tag{2.1}$$

where β and σ^2 are estimated parameters. However, as the data are measured at different locations, fitting this model may result in residual spatial correlation coming from unmeasured or unknown spatially dependent covariates that have not been included in the model. This can be accounted for by adding in spatial random effects.

Spatial effects can be represented in different ways. Here, we adopt a thin plate spline formulation which we implement in the `R`-package `mgcv` (for details of thin plate spline models, see Wahba [1990] and Chapter 5 of Wood [2017]). In this formulation, the spatial effect is modelled as an unknown smooth function defined on the spatial domain of the data. This function is designed to capture the spatial dependence of the residuals in the model (2.1) and is estimated using a criterion that aims to get as close to the data as possible, while simultaneously avoiding excessive function wiggliness, as such wiggliness is likely to lead to overfitting. In practice, this is done by a penalised version of maximum likelihood estimation where a penalty is imposed on the derivatives of the function. More specifically, we define the spatial model as

$$y_i = \beta x_i + f(\mathbf{t}_i) + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2). \tag{2.2}$$

where the estimates $\hat{\beta}$ and $\hat{f}$ are the minimisers of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta x_i - f(\mathbf{t}_i))^2 + \lambda \sum_{i_1,\ldots,i_m} \int_{\mathbb{R}^d} \left| \frac{\partial^m f(\mathbf{t})}{\partial t_{i_1} \cdots \partial t_{i_m}} \right|^2 d\mathbf{t}$$

with $\lambda > 0$ an unknown smoothing parameter (estimated from a separate criterion). Minimisation here is over all $\beta \in \mathbb{R}$ and functions $f \in H^m(\mathbb{R}^d)$ with $\frac{\partial^m f}{\partial t_{i_1} \cdots \partial t_{i_m}} \in L^2(\mathbb{R}^d)$ for

all subsets $i_1, \ldots, i_m$ of $1, \ldots, n$. Although this minimisation is over an infinite-dimensional space, as stated in Paper 1, the solution $\hat{f}$ lies in the n-dimensional space of thin plate splines, and the estimated effects $\hat{\beta}$ and $\widehat{\mathbf{f}} = (\hat{f}(\mathbf{t}_1), \ldots, \hat{f}(\mathbf{t}_n))^T$ can be found as the minimisers of

$$\|\mathbf{y} - \beta\mathbf{x} - \mathbf{f}\|^2 + n\lambda \mathbf{f}^T \boldsymbol{\Gamma} \mathbf{f} \tag{2.3}$$

with $\boldsymbol{\Gamma}$ an $n \times n$ penalty matrix. Solving the resulting normal equations, we see that

$$\hat{\beta} = \left(\mathbf{x}^T(\mathbf{I} - \mathbf{A}_\lambda)\mathbf{x}\right)^{-1} \mathbf{x}^T(\mathbf{I} - \mathbf{A}_\lambda)\mathbf{y}, \quad \widehat{\mathbf{f}} = \mathbf{A}_\lambda(\mathbf{y} - \hat{\beta}\mathbf{x}) \tag{2.4}$$

where $\mathbf{A}_\lambda = (\mathbf{I} + n\lambda\boldsymbol{\Gamma})^{-1}$ is known as the smoother matrix and is the influence matrix for the model (2.2) with no covariate term.

In Paper 1 we study the behaviour of these estimates assuming the covariate $\mathbf{x}$ has the form

$$x_i = f^x(\mathbf{t}_i) + \epsilon_i^x, \quad \epsilon_i^x \overset{\text{iid}}{\sim} N(0, \sigma_x^2) \tag{2.5}$$

where $f^x \in H^m(\Omega)$ is bounded. In other words, $\mathbf{x}$ decomposes into a smooth spatial part f^x and a non-spatial part $\boldsymbol{\epsilon}^x = (\epsilon_1^x, \ldots, \epsilon_n^x)^T$. We use asymptotic analysis (as the number of fitted data points $n \to \infty$) to theoretically show why bias in the effect estimate $\hat{\beta}$ occurs in this case. We also propose a method, spatial+, for avoiding this bias with asymptotic results to back it up. In the rest of this section, however, rather than a detailed asymptotic analysis, we provide some intuition for why spatial confounding problems occur in this case and why spatial+ works. In order to simplify notation, when referring to a matrix $\mathbf{M}$, we write $\mathbf{M}$ to denote both the matrix itself, the vectors making up the columns of $\mathbf{M}$ and the space spanned by the columns of $\mathbf{M}$.

From (2.3) we see that, using the above formulation, the spatial model is simply a linear model with model matrix $\mathbf{X} = [\mathbf{x}|\mathbf{B}_{\text{sp}}]$ where $\mathbf{B}_{\text{sp}}$ (which models the spatial effect $\mathbf{f}$) spans the space of thin plate spline functions defined on the spatial domain (evaluated at the data locations). The model is fitted through standard linear least squares but with a smoothing penalty applied to the spatial effect. In practice, in the `R`-package `mgcv`, we use a reduced rank approximation (known as thin plate regression splines) in which $\mathbf{B}_{\text{sp}}$ has a smaller number of columns. This number of columns, i.e. the number of basis functions used to represent the spatial effect, is determined by the user. The basis size should be chosen relatively small in order to reduce computation time for model fitting, but large enough to capture the spatial variation in the data. The basis is generally ordered with lower frequency spatial patterns first so that adding more spatial basis functions increases the ability of the spatial effect to model more complex spatial variation involving both lower and higher frequency spatial patterns. Typically, a relatively high number of basis functions is needed to allow sufficient flexibility in the spatial effect, but the smoothing penalty means that the effective degrees of freedom are reduced in order to avoid overfitting the data.

Using any basis $\mathbf{B}_{\text{sp}}$ for the spatial part of the model and writing $\mathbf{f} = \mathbf{B}_{\text{sp}}\boldsymbol{\beta}_{\text{sp}}$ in (2.3), the spatial model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{2.6}$$

with model matrix $\mathbf{X} = [\mathbf{x}|\mathbf{B}_{\text{sp}}]$ and where the unknown coefficients $\boldsymbol{\beta} = (\beta, \boldsymbol{\beta}_{\text{sp}}^T)^T$ are estimated as the minimisers of

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta} \tag{2.7}$$

with penalty matrix

$$\mathbf{S} = \begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{S}_{\text{sp}} \end{bmatrix}$$

where $\lambda \mathbf{B}_{\text{sp}} \mathbf{S}_{\text{sp}} \mathbf{B}_{\text{sp}}^T$ corresponds to $n\lambda\boldsymbol{\Gamma}$ in (2.3). Note that here, for ease of notation, we have absorbed n into the parameter λ. The normal equations then lead to the estimates

$$\begin{bmatrix} \hat{\beta} \\ \widehat{\boldsymbol{\beta}}_{\text{sp}} \end{bmatrix} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} \mathbf{x}^T\mathbf{x} & \mathbf{x}^T\mathbf{B}_{\text{sp}} \\ \mathbf{B}_{\text{sp}}^T\mathbf{x} & \mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x}^T \\ \mathbf{B}_{\text{sp}}^T \end{bmatrix} \mathbf{y}. \tag{2.8}$$

The resulting covariate effect estimate $\hat{\beta}$ and spatial effect estimate $\widehat{\mathbf{f}} = \mathbf{B}_{\text{sp}}\widehat{\boldsymbol{\beta}}_{\text{sp}}$ agree with the expressions in (2.4) but the smoother matrix $\mathbf{A}_\lambda$ now takes the form

$$\mathbf{A}_\lambda = \mathbf{B}_{\text{sp}}(\mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{sp}} + \lambda\mathbf{B}_{\text{sp}}\mathbf{S}_{\text{sp}}\mathbf{B}_{\text{sp}}^T)^{-1}\mathbf{B}_{\text{sp}}^T.$$

Note that the expression $\mathbf{A}_\lambda = (\mathbf{I} + n\lambda\boldsymbol{\Gamma})^{-1}$ for the smoother matrix assumes that we have used an orthonormal basis $\mathbf{B}_{\text{sp}}$ for the space of thin plate splines, which is convenient for the theoretical analysis in Paper 1.

In the following sections we use the model formulation (2.6) to analyse the behaviour of effect estimates in the spatial model. Initially, in Section 2.2, we consider the model without the smoothing penalty applied (i.e. where $\lambda = 0$ in (2.7)) and, in Section 2.3, we then consider the effects of smoothing. Throughout our analysis we assume that the covariate $\mathbf{x}$ is not contained in the span of the spatial basis $\mathbf{B}_{\text{sp}}$, i.e. while $\mathbf{x}$ may be spatially dependent, it is not fully determined by spatial location. We show that the non-spatial part of the covariate can be used to obtain unbiased covariate effect estimates in a relatively simple way. In contrast, if $\mathbf{x}$ lies in $\mathbf{B}_{\text{sp}}$, the model matrix of the spatial model has linearly dependent columns and, thus, without the smoothing penalty, the model is unidentifiable. In practical terms, the unidentifiable model cannot distinguish the effect of $\mathbf{x}$ from that of its spatial pattern, which could be shared by other covariates or caused by other underlying spatial processes. The apportionment of the total effect of this spatial pattern on the response variable between the covariate and the spatial terms in the model is, in this case, entirely determined by the smoothing penalty, and this can lead to rather arbitrary results. The resulting bias in the covariate effect estimate has been studied by Paciorek [2010] and Page et al. [2017] who show that the size of the bias depends on the relative spatial scales of the covariate and the spatial effects. Separating out the true effects in this situation is more complicated and the problem is typically analysed in the context of causal inference (an overview of the literature can be found in Reich et al. [2020]).

Estimates without smoothing

If, in the first instance, we ignore the smoothing penalty in (2.7), we see that the spatial model (2.6) is an ordinary linear model with model matrix $\mathbf{X} = [\mathbf{x}|\mathbf{B}_{\text{sp}}]$ and parameters $\boldsymbol{\beta} = (\beta, \boldsymbol{\beta}_{\text{sp}}^T)^T$ estimated through linear least squares. In particular, the covariate effect estimate $\hat{\beta}$ is unbiased. This shows that, in the absence of smoothing, the spatial model is actually able to capture the true effect. This is perhaps surprising as the common perception, as well as the motivation behind RSR, seems to be that the estimate in the null model (2.1) is "correct" and should be preserved. This perception, however, is not in general true which can be seen as follows.

Suppose the response data is actually generated as

$$\mathbf{y} = \beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

so that β is the true covariate effect and $\mathbf{f} = \mathbf{B}_{\text{sp}}\boldsymbol{\beta}_{\text{sp}}$ the true unmeasured spatial effect. The effect estimate $\hat{\beta}_{\text{null}}$ in the null model is then given by

$$\begin{aligned}\hat{\beta}_{\text{null}} &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \\ &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T(\beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}) \\ &= \beta + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{f} + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\boldsymbol{\epsilon}\end{aligned}$$

so that

$$\mathrm{E}(\hat{\beta}_{\text{null}}) = \beta + \mathrm{E}\big[(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{f}\big].$$

So unless $\mathbf{x}$ is independent of the true spatial effect $\mathbf{f}$, the estimate in the null model is biased, and the more correlated $\mathbf{x}$ is with $\mathbf{f}$, the larger the bias. This is because the model matrix in the null model consists of the single column $\mathbf{x}$ and has no other component to explain the part of $\mathbf{y}$ that is not iid noise. Thus, the covariate term will reflect, not only the effect of $\mathbf{x}$, but also any part of the true spatial effect $\mathbf{f}$ that is similar to $\mathbf{x}$. In other words, $\mathbf{x}$ acts as a proxy for any unmeasured covariates with a similar spatial pattern and, therefore, we do not recover the true effect of $\mathbf{x}$ in our estimate.

In contrast, the model matrix of the spatial model (2.6) includes the spatial part $\mathbf{B}_{\text{sp}}$ for explaining the spatial effect $\mathbf{f}$. In fact, the model generating the data is the same linear model as the spatial model fitted to it, so perhaps it is no surprise that the true effects are recaptured in the estimation. More precisely, the estimates in the unsmoothed spatial model are given by

$$\begin{bmatrix}\hat{\beta} \\ \widehat{\boldsymbol{\beta}}_{\text{sp}}\end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix}\mathbf{x}^T\mathbf{x} & \mathbf{x}^T\mathbf{B}_{\text{sp}} \\ \mathbf{B}_{\text{sp}}^T\mathbf{x} & \mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{sp}}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{x}^T \\ \mathbf{B}_{\text{sp}}^T\end{bmatrix}\mathbf{y}.$$

Though a formula exists for the above matrix inversion, this direct approach to calculating the estimate of interest $\hat{\beta}$ is somewhat tedious and not necessarily illuminating at this point. Instead we use a reparametrisation that simplifies the calculation of $\hat{\beta}$ and also motivates the spatial+ approach which we propose in Paper 1 and briefly discuss in Section 2.3.

Let $\mathbf{P}_{\text{sp}} : \mathbb{R}^n \to \mathbb{R}^n$ be the orthogonal projection onto $\mathbf{B}_{\text{sp}}$, i.e.

$$\mathbf{P}_{\text{sp}} = \mathbf{B}_{\text{sp}}(\mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{sp}})^T\mathbf{B}_{\text{sp}}^T.$$

Then $\mathbf{x}$ decomposes as

$$\mathbf{x} = \mathbf{P}_{\text{sp}}\mathbf{x} + \mathbf{r}$$

where $\mathbf{P}_{\text{sp}}\mathbf{x}$ lies in $\mathbf{B}_{\text{sp}}$ and $\mathbf{r} = (\mathbf{I} - \mathbf{P}_{\text{sp}})\mathbf{x}$ is the projection of $\mathbf{x}$ onto the orthogonal complement of $\mathbf{B}_{\text{sp}}$. Replacing $\mathbf{X} = [\mathbf{x}|\mathbf{B}_{\text{sp}}]$ by $\mathbf{X}' = [\mathbf{r}|\mathbf{B}_{\text{sp}}]$ we see that $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\boldsymbol{\beta}'$ where $\boldsymbol{\beta}' = (\beta, \boldsymbol{\beta}_{\text{sp}}'^T)^T$ with $\boldsymbol{\beta}_{\text{sp}}' = \beta(\mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{sp}})^{-1}\mathbf{B}_{\text{sp}}^T\mathbf{x} + \boldsymbol{\beta}_{\text{sp}}$. In other words, reparametrising the spatial model as

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta}' + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{2.9}$$

will result in the same fitted values as the spatial model and the effect of $\mathbf{r}$ in this model is

the same as the effect of $\mathbf{x}$ in the spatial model. Since $\mathbf{r}^T\mathbf{B}_{\text{sp}} = \mathbf{0}$, we see that

$$\begin{bmatrix} \hat{\beta} \\ \widehat{\boldsymbol{\beta}}'_{\text{sp}} \end{bmatrix} = (\mathbf{X}'^T\mathbf{X}')^{-1}\mathbf{X}'^T\mathbf{y} = \begin{bmatrix} (\mathbf{r}^T\mathbf{r})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{sp}})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{r}^T \\ \mathbf{B}_{\text{sp}}^T \end{bmatrix} \mathbf{y},$$

in particular,

$$\begin{aligned} \hat{\beta} &= (\mathbf{r}^T\mathbf{r})^{-1}\mathbf{r}^T\mathbf{y} \\ &= (\mathbf{r}^T\mathbf{r})^{-1}\mathbf{r}^T(\beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}) \\ &= \beta + (\mathbf{r}^T\mathbf{r})^{-1}\mathbf{r}^T\boldsymbol{\epsilon} \end{aligned}$$

and hence

$$\mathrm{E}(\hat{\beta}) = \beta.$$

The reparametrisation illustrates two things. Firstly, the estimated effect $\hat{\beta}$ in the unsmoothed spatial model is unbiased as expected. Secondly, $\mathbf{r}$, the part of $\mathbf{x}$ that is orthogonal to $\mathbf{B}_{\text{sp}}$, contains all the information needed to capture the effect of $\mathbf{x}$ on $\mathbf{y}$. Thus, any part of $\mathbf{x}$ that can be explained by the spatial basis vectors $\mathbf{B}_{\text{sp}}$ is unnecessary for estimating β. More precisely, for any vector $\mathbf{v}$ in $\mathbf{B}_{\text{sp}}$, replacing $\mathbf{x}$ by $\mathbf{x} - \mathbf{v}$ in the unsmoothed spatial model makes no difference to the covariate effect estimate $\hat{\beta}$ or the fitted values of the model.

It is generally assumed that collinearity between the covariate and spatial effects in the spatial model is the underlying cause of spatial confounding bias. Here, we use the word collinear in a broad sense to mean that $\mathbf{x}^T\mathbf{B}_{\text{sp}} \neq \mathbf{0}$, i.e. that $\mathbf{x}$ is associated with the spatial basis functions in some way. Our analysis shows that, even if $\mathbf{x}$ and $\mathbf{B}_{\text{sp}}$ are highly collinear, since we have assumed that the model is identifiable (i.e. that $\mathbf{x}$ is not contained in $\mathbf{B}_{\text{sp}}$), the estimated effect $\hat{\beta}$ is still unbiased. This shows that spatial confounding in this context cannot be fully understood without also considering the effects of smoothing, i.e. of imposing a penalty on the wiggliness of the spatial effect.

Estimates with smoothing

Smoothing is introduced in the spatial model in order to achieve an overall better fit to the data. We allow bias in the estimates but in return for lower variance of fitted values, and we estimate the smoothing parameter λ with a view to obtaining an optimal balance in this bias-variance trade-off. Different smoothness selection criteria are briefly described in Paper 1 and more details can be found in Wood [2017]. Although the smoothing penalty is only applied to the spatial part of the model, if $\mathbf{x}$ is collinear with $\mathbf{B}_{\text{sp}}$, the estimated effect $\hat{\beta}$ is also affected by the smoothing penalty even though it relates to the unpenalised part of the model. This is perhaps surprising but can be seen by looking more closely at the model estimates (2.8).

From e.g. Wood [2017] p.80, we have that for square matrices $\mathbf{A}$ and $\mathbf{B}$

$$\begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}\mathbf{D}^{-1}\mathbf{C}^T\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{C}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}^T\mathbf{A}^{-1} & \mathbf{D}^{-1} \end{bmatrix}$$

if $\mathbf{A}$ and $\mathbf{D} = \mathbf{B} - \mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}$ are invertible. Using this with $\mathbf{A} = \mathbf{x}^T\mathbf{x}$, $\mathbf{B} = \mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}}$,

$\mathbf{C} = \mathbf{x}^T\mathbf{B}_{\text{sp}}$ and

$$\begin{aligned}\mathbf{D} &= \mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}} - \mathbf{B}_{\text{sp}}^T\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{B}_{\text{sp}} \\ &= \mathbf{B}_{\text{sp}}^T(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}},\end{aligned}$$

we see that the estimated spatial effect $\widehat{\mathbf{f}} = \mathbf{B}_{\text{sp}}\widehat{\boldsymbol{\beta}}_{\text{sp}}$ is given by

$$\begin{aligned}\widehat{\mathbf{f}} = \mathbf{B}_{\text{sp}}\widehat{\boldsymbol{\beta}}_{\text{sp}} &= \mathbf{B}_{\text{sp}}(-\mathbf{D}^{-1}\mathbf{C}^T\mathbf{A}^{-1}\mathbf{x}^T\mathbf{y} + \mathbf{D}^{-1}\mathbf{B}_{\text{sp}}^T\mathbf{y}) \\ &= \mathbf{B}_{\text{sp}}\mathbf{D}^{-1}\big(\mathbf{B}_{\text{sp}}^T\mathbf{y} - \mathbf{C}^T(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}\big) \\ &= \mathbf{B}_{\text{sp}}\big(\mathbf{B}_{\text{sp}}^T(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}}\big)^{-1}\big(\mathbf{B}_{\text{sp}}^T\mathbf{y} - \mathbf{B}_{\text{sp}}^T(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}\big) \\ &= \mathbf{B}_{\text{sp}}\big(\mathbf{B}_{\text{sp}}^T(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}}\big)^{-1}\mathbf{B}_{\text{sp}}^T(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{y}\end{aligned} \tag{2.10}$$

and

$$\begin{aligned}\hat{\beta} &= (\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}\mathbf{D}^{-1}\mathbf{C}^T\mathbf{A}^{-1})\mathbf{x}^T\mathbf{y} - \mathbf{A}^{-1}\mathbf{C}\mathbf{D}^{-1}\mathbf{B}_{\text{sp}}^T\mathbf{y} \\ &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \\ &\quad +(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{B}_{\text{sp}}\big(\mathbf{B}_{\text{sp}}^T(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}}\big)^{-1}\mathbf{B}_{\text{sp}}^T\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \\ &\quad -(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{B}_{\text{sp}}\big(\mathbf{B}_{\text{sp}}^T(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}}\big)^{-1}\mathbf{B}_{\text{sp}}^T\mathbf{y} \\ &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T(\mathbf{y} - \widehat{\mathbf{f}}).\end{aligned}$$

Assuming, as we did before, that $\mathbf{y}$ has the form $\mathbf{y} = \beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}$ so that β and $\mathbf{f}$ are the true covariate and spatial effects, respectively, we therefore see that

$$\begin{aligned}\text{E}(\hat{\beta}) &= \text{E}\big[(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T(\beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon} - \widehat{\mathbf{f}})\big] \\ &= \beta + \text{E}\big[(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T(\mathbf{f} - \widehat{\mathbf{f}})\big].\end{aligned}$$

The above expressions show that, unsurprisingly, the estimated spatial effect $\widehat{\mathbf{f}}$ depends directly on the smoothing penalty $\lambda\mathbf{S}_{\text{sp}}$. We know that this penalty induces bias in $\widehat{\mathbf{f}}$ as $\lambda\mathbf{S}_{\text{sp}} = \mathbf{0}$ corresponds to the unbiased estimate obtained in the unsmoothed spatial model analysed in Section 2.2. But now, since $\hat{\beta}$ depends on $\widehat{\mathbf{f}}$, it will also be affected by smoothing even though the parameter β does not relate to the penalised part of the model. In other words, by allowing bias in the spatial effect estimate, we indirectly allow bias in the covariate effect estimate as well. Looking at the expression for $\text{E}(\hat{\beta})$, we see that this bias depends on two things: how close the estimate $\widehat{\mathbf{f}}$ is to the true effect $\mathbf{f}$ (which in turn depends on how much smoothing is applied); and how correlated $\mathbf{x}$ is with the spatial basis vectors (since $\mathbf{f} - \widehat{\mathbf{f}} = \mathbf{B}_{\text{sp}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})$ lies in $\mathbf{B}_{\text{sp}}$). In fact the expression shows that, although collinearity in itself (without smoothing) does not cause bias in the estimate $\hat{\beta}$, if $\mathbf{x}$ is highly collinear with $\mathbf{B}_{\text{sp}}$, then $\hat{\beta}$ may be sensitive to changes in the estimate $\widehat{\mathbf{f}}$, in particular, to the effects of smoothing. Thus, spatial confounding arises as a result of the combined effect of collinearity and smoothing.

This understanding is key to the idea behind the spatial+ model which we propose as a method for dealing with spatial confounding in Paper 1. Collinearity between $\mathbf{x}$ and $\mathbf{B}_{\text{sp}}$ is what makes the estimate of β in the spatial model biased when smoothing is applied. This collinearity is exactly caused by the spatial dependence of $\mathbf{x}$. However, as the reparametrisation (2.9) of the unsmoothed spatial model shows, the spatial part of $\mathbf{x}$ is

unnecessary for capturing the true effect β, and we can replace $\mathbf{x}$ in the model matrix of this model by $\mathbf{x} - \mathbf{v}$ for any vector $\mathbf{v}$ in $\mathbf{B}_{\text{sp}}$ without altering the estimate of β or the fitted values. The spatial+ model is a smoothed version of such a reparametrisation where $\mathbf{v}$ is the estimated spatial pattern of $\mathbf{x}$.

More specifically, the model is defined in two steps. First, a spatial model is fitted to the covariate $\mathbf{x}$, i.e.

$$x_i = f^x(\mathbf{t}_i) + \epsilon_i^x, \quad \epsilon_i^x \underset{\text{iid}}{\sim} N(0, \sigma_x^2). \tag{2.11}$$

where f^x is a thin plate spline. Thus, the fitted values $\widehat{\mathbf{f}}^x = (\hat{f}^x(\mathbf{t}_1), \ldots, \hat{f}^x(\mathbf{t}_n))^T$ in this model represent the spatial pattern of $\mathbf{x}$, and we obtain the decomposition

$$\mathbf{x} = \widehat{\mathbf{f}}^x + \mathbf{r}^x$$

where $\mathbf{r}^x = (r_1^x, \ldots, r_n^x)^T$ are the residuals in the model. The spatial+ model is then obtained by replacing the covariate $\mathbf{x}$ in the spatial model (2.2) by $\mathbf{r}^x$, i.e.

$$y_i = \beta r_i^x + f^+(\mathbf{t}_i) + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2)$$

with β and f^+ estimated as before. With appropriate adjustments to the above derivations, we see that the estimated covariate effect $\hat{\beta}^+$ in this model is given by

$$\hat{\beta}^+ = \left(\mathbf{r}^{xT}\mathbf{r}^x\right)^{-1}\mathbf{r}^{xT}(\mathbf{y} - \widehat{\mathbf{f}^+})$$

and, once again assuming the data has true effects β and $\mathbf{f}$, we have that

$$\mathbf{y} = \beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon} = \beta\mathbf{r}^x + \mathbf{f}^+ + \boldsymbol{\epsilon}$$

where $\mathbf{f}^+ = \beta\widehat{\mathbf{f}}^x + \mathbf{f}$ and, therefore,

$$\mathrm{E}(\hat{\beta}^+) = \beta + \mathrm{E}\big[(\mathbf{r}^{xT}\mathbf{r}^x)^{-1}\mathbf{r}^{xT}(\mathbf{f}^+ - \widehat{\mathbf{f}^+})\big].$$

Thus, although the estimate $\hat{\beta}^+$ is biased, we would expect the size of the bias to be relatively small as the residuals $\mathbf{r}^x$ and the term $\mathbf{f}^+ - \widehat{\mathbf{f}^+}$ would have only little collinearity. This is because the residuals $\mathbf{r}^x$ are exactly the part of $\mathbf{x}$ that cannot be explained by spatial location and as such would be largely orthogonal to $\mathbf{B}_{\text{sp}}$ which contains $\mathbf{f}^+ - \widehat{\mathbf{f}^+}$. In other words, by largely decoupling the estimation of the covariate effect β from the spatial part of the model, the estimate becomes relatively unaffected by smoothing and remains close to the unbiased estimate in the unsmoothed model.

Non-Gaussian response distributions

As also noted in Paper 1, the formulation (2.2) of the spatial model generalises to non-Gaussian response distributions, specifically, those from the exponential family of distributions. A distribution is in this family if its probability density function p can be written in the form

$$p(y) = \exp\left[\{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)\right]$$

where θ and ϕ are parameters of the distribution and a, b and c are functions. Examples include the Gaussian, Poisson, gamma and binomial distributions. Suppose the response data $\mathbf{y} = (y_1, \ldots, y_n)^T$ are observations of random variables y_i from the exponential family of distributions with $\mathrm{E}(y_i) = \mu_i$, and suppose $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{t}_1, \ldots, \mathbf{t}_n$ are covariate observations and spatial locations as before. A generalised version of (2.2) can then be formulated as

$$g(\mu_i) = \beta x_i + f(\mathbf{t}_i) \tag{2.12}$$

where β is an unknown parameter, f a thin plate spline and $g : \mathbb{R} \to \mathbb{R}$ a link function (i.e. a monotonic smooth function which ensures $g(\mu_i)$ is in the domain of the response variable).

Using the same setup as (2.6), i.e. the model matrix $\mathbf{X} = [\mathbf{x}|\mathbf{B}_{\mathrm{sp}}]$ with spatial basis $\mathbf{B}_{\mathrm{sp}}$ and unknown coefficients $\boldsymbol{\beta} = (\beta, \boldsymbol{\beta}_{\mathrm{sp}}^T)^T$, the estimated effects $\hat{\beta}$ and $\widehat{\mathbf{f}} = \mathbf{B}_{\mathrm{sp}}\widehat{\boldsymbol{\beta}}_{\mathrm{sp}}$ are obtained using an iterative version of penalised least squares known as the penalised iterative re-weighted least squares (PIRLS) algorithm, described e.g. in Wood [2017]. Starting at step $k = 1$ with initial values for $\hat{\mu}_i^{[k]}$, $i = 1, \ldots, n$, the algorithm is iterated by replacing $\hat{\mu}_i^{[k]}$ by $\hat{\mu}_i^{[k+1]} = g^{-1}(\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{[k+1]})$ where $\widehat{\boldsymbol{\beta}}^{[k+1]}$ is the minimiser of

$$\|\sqrt{\mathbf{W}^{[k]}}(\mathbf{z}^{[k]} - \mathbf{X}\boldsymbol{\beta})\|^2 + \phi\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta}$$

with $\mathbf{S}$ the smoothing penalty from before and $\mathbf{W}^{[k]} = \mathrm{diag}(w_1^{[k]}, \ldots, w_n^{[k]})$ and $\mathbf{z}^{[k]} = (z_1^{[k]}, \ldots, z_n^{[k]})^T$ the weights matrix and pseudodata which are given by

$$w_i^{[k]} = 1/(g'(\hat{\mu}_i^{[k]})^2 V(\hat{\mu}_i^{[k]})), \quad z_i^{[k]} = g'(\hat{\mu}_i^{[k]})(y_i - \hat{\mu}_i^{[k]}) + g(\hat{\mu}_i^{[k]})$$

with V the variance function for the response distribution. Repeating this process until convergence leads to the estimate

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X} + \phi\mathbf{S})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z} \tag{2.13}$$

where $\mathbf{W}$ and $\mathbf{z}$ are the weights and pseudodata at convergence which depend on $\hat{\mu}_i = g^{-1}(\mathbf{X}_i\widehat{\boldsymbol{\beta}})$, the fitted values of the model. If no smoothing penalty is applied, the estimate $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimate in a generalized linear model (GLM). While the GLM estimate is in general no longer unbiased as it was in the Gaussian case, it is asymptotically unbiased as the number of fitted data points $n \to \infty$.

Note that in the final step of the PIRLS algorithm, the estimate $\widehat{\boldsymbol{\beta}}$ is obtained through the minimisation of (2.7) but where we replace $\mathbf{y}$, $\mathbf{x}$, $\mathbf{B}_{\mathrm{sp}}$ and $\mathbf{S}$ by

$$\check{\mathbf{y}} = \sqrt{\mathbf{W}}\mathbf{z}, \quad \check{\mathbf{x}} = \sqrt{\mathbf{W}}\mathbf{x}, \quad \check{\mathbf{B}}_{\mathrm{sp}} = \sqrt{\mathbf{W}}\mathbf{B}_{\mathrm{sp}}, \quad \check{\mathbf{S}} = \phi\mathbf{S}. \tag{2.14}$$

Thus, the estimate corresponds to that of a Gaussian spatial model with model matrix $\check{\mathbf{X}} = [\check{\mathbf{x}}|\check{\mathbf{B}}_{\mathrm{sp}}]$ and smoothing penalty $\check{\mathbf{S}}$. Hence, our analysis of the Gaussian case can be used to understand the behaviour of the generalised model as well. In particular, we can substitute (2.14) into the expressions for $\widehat{\boldsymbol{\beta}}_{\mathrm{sp}}$ and $\hat{\beta}$ in Section 2.3 to obtain the corresponding estimates for the model (2.12). Using this, we can show that

$$\begin{aligned} \hat{\beta} &= (\check{\mathbf{x}}^T\check{\mathbf{x}})^{-1}\check{\mathbf{x}}^T(\check{\mathbf{y}} - \sqrt{\mathbf{W}}\widehat{\mathbf{f}}) \\ &= (\mathbf{x}^T\mathbf{W}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}(\mathbf{z} - \widehat{\mathbf{f}}) \end{aligned}$$

where $\widehat{\mathbf{f}}$ depends on the smoothing penalty. So once again the covariate effect estimate $\hat{\beta}$, which is asymptotically unbiased in the absence of smoothing, becomes biased when smoothing is applied due to its dependence on the estimated spatial effect. As before, the sensitivity to this is determined by the collinearity between $\mathbf{x}$ and $\mathbf{B}_{\text{sp}}$, however, now collinearity is measured by $\mathbf{x}^T\mathbf{W}\mathbf{B}_{\text{sp}}$, i.e. in terms of the inner product defined by the weights matrix $\mathbf{W}$.

As shown in Paper 1, a non-Gaussian version of the spatial+ model can also be defined. Let $\widehat{\mathbf{f}}^x$ and $\mathbf{r}^x = \mathbf{x} - \widehat{\mathbf{f}}^x = (r_1^x, \ldots, r_n^x)^T$ denote the fitted values and residuals in the weighted version of the thin plate spline regression (2.11) with weights $\mathbf{W}$, i.e. $\widehat{\mathbf{f}}^x = \mathbf{B}_{\text{sp}}\widehat{\boldsymbol{\beta}}^x_{\text{sp}}$ where $\widehat{\boldsymbol{\beta}}^x_{\text{sp}}$ is the minimiser of

$$\|\sqrt{\mathbf{W}}(\mathbf{x} - \mathbf{B}_{\text{sp}}\boldsymbol{\beta}^x_{\text{sp}})\|^2 + \lambda_x \widehat{\boldsymbol{\beta}}^{xT}_{\text{sp}}\mathbf{S}_{\text{sp}}\widehat{\boldsymbol{\beta}}^x_{\text{sp}}$$

with smoothing parameter $\lambda_x > 0$. By the properties of weighted thin plate spline regressions $\mathbf{r}^{xT}\mathbf{W}\mathbf{B}_{\text{sp}} \approx \mathbf{0}$. The spatial+ model is then the spatial model (2.12) with the covariate $\mathbf{x}$ replaced by $\mathbf{r}^x$, i.e.

$$g(\mu_i) = \beta r_i^x + f^+(\mathbf{t}_i) \tag{2.15}$$

where β and f^+ are estimated as before. In the Gaussian formulation with substitutions (2.14), i.e. a spatial model with model matrix $\check{\mathbf{X}} = [\check{\mathbf{x}}|\check{\mathbf{B}}_{\text{sp}}]$ and smoothing penalty $\check{\mathbf{S}}$, the spatial+ model therefore corresponds to replacing $\check{\mathbf{x}}$ by $\check{\mathbf{r}}^x = \sqrt{\mathbf{W}}\mathbf{r}^x$ in the model matrix. Since

$$\check{\mathbf{x}} = \check{\mathbf{r}}^x + \sqrt{\mathbf{W}}\widehat{\mathbf{f}}^x$$

is a decomposition in which $\sqrt{\mathbf{W}}\widehat{\mathbf{f}}^x$ lies in $\check{\mathbf{B}}_{\text{sp}}$, the component $\check{\mathbf{r}}^x$ has the same effect β on the response as $\check{\mathbf{x}}$, but $\check{\mathbf{r}}^x$ is broadly orthogonal to $\check{\mathbf{B}}_{\text{sp}} = \sqrt{\mathbf{W}}\mathbf{B}_{\text{sp}}$ as $\mathbf{r}^x\mathbf{W}\mathbf{B}_{\text{sp}} \approx \mathbf{0}$. Thus, the estimation of covariate and spatial effects are largely decoupled and the estimate of β stays broadly in line with the asymptotically unbiased estimate in the unsmoothed spatial model when the smoothing penalty is applied.

RSR

RSR, i.e. the method of orthogonalisation applied to spatial models, was first introduced by Reich et al. [2006] for discrete space models in which spatial correlation is modelled by an intrinsic conditional autoregressive (ICAR) random effect. The method was extended to continuous space models by Hanks et al. [2015] and is commonly used to avoid spatial confounding bias. The idea behind this method is to restrict the spatial effects in the spatial model to the orthogonal complement of the covariate so that they cannot interfere with the covariate effect estimates. However, our analysis here shows that this is not an effective way of eliminating spatial confounding bias, which is also confirmed in recent papers [Khan and Calder, 2020, Nobre et al., 2020].

Consider a linear model with response variable $\mathbf{y}$ and two observed covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ which are highly collinear. The model would, in this case, be poorly identifiable and may have difficulty correctly apportioning out the variability in $\mathbf{y}$ between the two variables. Let $\widetilde{\mathbf{x}}_2$ be the projection of $\mathbf{x}_2$ onto the orthogonal complement of $\mathbf{x}_1$, i.e.

$$\widetilde{\mathbf{x}}_2 = (\mathbf{I} - \mathbf{x}_1(\mathbf{x}_1^T\mathbf{x}_1)^{-1}\mathbf{x}_1^T)\mathbf{x}_2.$$

The orthogonalised model is obtained by replacing $\mathbf{x}_2$ by $\widetilde{\mathbf{x}}_2$ in the model matrix. Thus, the column space of the model matrix, and hence the fitted values, remains the same as in

the original model but, by construction, $\mathbf{x}_1$ and $\widetilde{\mathbf{x}}_2$ are independent. This means that all variability in $\mathbf{y}$ that could previously have been explained by either of the original covariates is now attributed to $\mathbf{x}_1$ and the variable $\widetilde{\mathbf{x}}_2$ can only explain the residual variability. In a general setting, this prioritisation of one variable over another can be problematic as it can be difficult to know which covariate is the more important. But in the spatial model, Reich et al. [2006] argue that, since the spatial effects can be viewed as simply a technical tool used to improve model-fitting, it seems more natural to restrict the spatial effects to only explain spatial variation that cannot be attributed to any observed covariates.

In the setting of Sections 2.2 and 2.3, RSR is defined as follows. Let $\mathbf{P} : \mathbb{R}^n \to \mathbb{R}^n$ denote the projection onto the span of the covariate $\mathbf{x}$, i.e.

$$\mathbf{P} = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$$

so that

$$\mathbf{P}^c = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$$

is the projection onto the orthogonal complement of $\mathbf{x}$, and let

$$\widetilde{\mathbf{B}}_{\text{sp}} = \mathbf{P}^c\mathbf{B}_{\text{sp}}.$$

The RSR model is then defined in the same way as the spatial model (2.6) but with the model matrix $\mathbf{X} = [\mathbf{x}|\mathbf{B}_{\text{sp}}]$ replaced by $\widetilde{\mathbf{X}} = [\mathbf{x}|\widetilde{\mathbf{B}}_{\text{sp}}]$, i.e.

$$\mathbf{y} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

with unknown coefficients $\widetilde{\boldsymbol{\beta}} = (\tilde{\beta}, \widetilde{\boldsymbol{\beta}}_{\text{sp}}^T)^T$. As the column space of $\widetilde{\mathbf{X}}$ is the same as that of $\mathbf{X}$, we would expect similar fitted values to the spatial model (the only difference being that the estimated smoothing parameter may be slightly different). But now, as $\mathbf{x}^T\widetilde{\mathbf{B}}_{\text{sp}} = \mathbf{0}$, the estimated spatial effect $\widehat{\widetilde{\mathbf{f}}} = \widetilde{\mathbf{B}}_{\text{sp}}\widehat{\widetilde{\boldsymbol{\beta}}}_{\text{sp}}$ is orthogonal to $\mathbf{x}$, and the estimates in this model are given by

$$\begin{bmatrix} \hat{\tilde{\beta}} \\ \widehat{\widetilde{\boldsymbol{\beta}}}_{\text{sp}} \end{bmatrix} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} + \lambda\mathbf{S})^{-1}\widetilde{\mathbf{X}}^T\mathbf{y} = \begin{bmatrix} (\mathbf{x}^T\mathbf{x})^{-1} & \mathbf{0} \\ \mathbf{0} & (\widetilde{\mathbf{B}}_{\text{sp}}^T\widetilde{\mathbf{B}}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x}^T \\ \widetilde{\mathbf{B}}_{\text{sp}}^T \end{bmatrix} \mathbf{y}. \tag{2.16}$$

Thus, the estimated covariate effect is

$$\hat{\tilde{\beta}} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \tag{2.17}$$

which is the same as the estimate in the null model. As $\hat{\tilde{\beta}}$ is independent of the spatial effect estimate, it stays constant under spatial smoothing. However, as we have shown in Section 2.2, this estimate is biased unless the covariate $\mathbf{x}$ is independent of the true unmeasured spatial effect. Hence, although RSR makes the covariate effect estimate independent of smoothing, it is not an effective method for eliminating spatial confounding bias.

Another issue with RSR is the interpretation of the estimated spatial effect. We see from the above that this estimate is given by

$$\widehat{\widetilde{\mathbf{f}}} = \widetilde{\mathbf{B}}_{\text{sp}}\widehat{\widetilde{\boldsymbol{\beta}}}_{\text{sp}} = \widetilde{\mathbf{B}}_{\text{sp}}(\widetilde{\mathbf{B}}_{\text{sp}}^T\widetilde{\mathbf{B}}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}})^{-1}\widetilde{\mathbf{B}}_{\text{sp}}^T\mathbf{y}$$

and lies in the space $\widetilde{\mathbf{B}}_{\text{sp}}$ by construction. Note that, since we have assumed that $\mathbf{x}$ is not fully spatial (i.e. that $\mathbf{x}$ is not in $\mathbf{B}_{\text{sp}}$), the restriction $\mathbf{P}^c|_{\mathbf{B}_{\text{sp}}} : \mathbf{B}_{\text{sp}} \to \widetilde{\mathbf{B}}_{\text{sp}}$ of $\mathbf{P}^c$ to $\mathbf{B}_{\text{sp}}$ is an invertible map between spaces of the same dimension. Thus, although the name RSR may suggest that the orthogonalised spatial effect is restricted to a smaller space, in fact, the orthogonalisation simply moves the estimation into a different subspace of the same dimension as $\mathbf{B}_{\text{sp}}$. Looking at the expression (2.10) for the estimated spatial effect $\widehat{\mathbf{f}}$ in the spatial model, we see that

$$\begin{aligned}
\widehat{\mathbf{f}} &= \mathbf{B}_{\text{sp}}\big(\mathbf{B}_{\text{sp}}^T(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}}\big)^{-1}\mathbf{B}_{\text{sp}}^T(\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{y} \\
&= \mathbf{B}_{\text{sp}}\big(\mathbf{B}_{\text{sp}}^T\mathbf{P}^c\mathbf{B}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}}\big)^{-1}\mathbf{B}_{\text{sp}}^T\mathbf{P}^c\mathbf{y} \\
&= \mathbf{B}_{\text{sp}}\big(\widetilde{\mathbf{B}}_{\text{sp}}^T\widetilde{\mathbf{B}}_{\text{sp}} + \lambda\mathbf{S}_{\text{sp}}\big)^{-1}\widetilde{\mathbf{B}}_{\text{sp}}^T\mathbf{y}.
\end{aligned}$$

Hence, for the same value of the smoothing parameter λ, the estimated spatial effect in the RSR model is given by

$$\widehat{\widetilde{\mathbf{f}}} = \mathbf{P}^c\widehat{\mathbf{f}}.$$

Now, like any vector in $\mathbf{B}_{\text{sp}}$, $\widehat{\mathbf{f}}$ can be decomposed as

$$\widehat{\mathbf{f}} = \mathbf{P}\widehat{\mathbf{f}} + \mathbf{P}^c\widehat{\mathbf{f}}$$

with $\mathbf{P}\widehat{\mathbf{f}}$ in the span of $\mathbf{x}$ and $\mathbf{P}^c\widehat{\mathbf{f}}$ in $\widetilde{\mathbf{B}}_{\text{sp}}$ as illustrated in Figure 2-1. So in this sense, the RSR estimate $\widehat{\widetilde{\mathbf{f}}}$ can be interpreted as the component of the spatial model estimate $\widehat{\mathbf{f}}$ that is orthogonal to $\mathbf{x}$. However, since we have assumed that $\mathbf{x}$ is not in $\mathbf{B}_{\text{sp}}$, the component $\mathbf{P}\widehat{\mathbf{f}}$ does not lie in $\mathbf{B}_{\text{sp}}$ (unless it is zero) and therefore, the only way the component $\mathbf{P}^c\widehat{\mathbf{f}}$ can lie in $\mathbf{B}_{\text{sp}}$ is if $\mathbf{P}\widehat{\mathbf{f}} = \mathbf{0}$ and $\widehat{\mathbf{f}} = \mathbf{P}^c\widehat{\mathbf{f}}$. This shows that, unless the estimated spatial effect $\widehat{\mathbf{f}}$ is itself orthogonal to the covariate (such that the estimates in the spatial model and RSR model agree), then the RSR estimate is actually not "spatial" as it does not lie in $\mathbf{B}_{\text{sp}}$. This could explain why, for example, Hughes and Haran [2013] observe "non-spatial" behaviour in estimated spatial effects in their RSR model. Figure 2-1 also shows that the estimate $\widehat{\widetilde{\mathbf{f}}}$ could be quite far from (and much smaller than) the true residual spatial effect (which is captured by $\widehat{\mathbf{f}}$ if smoothing is ignored), especially if the covariate $\mathbf{x}$ is very spatially dependent.

RSR is also used for spatial models with non-Gaussian response distributions. Indeed, in the paper Reich et al. [2006] which introduces the technique, it is applied to a model with Poisson-distributed response data. In our formulation (2.12) of the generalised spatial model, orthogonalisation is achieved by replacing $\mathbf{B}_{\text{sp}}$ in the model matrix $\mathbf{X} = [\mathbf{x}|\mathbf{B}_{\text{sp}}]$ by $\widetilde{\mathbf{B}}_{\text{sp}} = \mathbf{P}^c\mathbf{B}_{\text{sp}}$ where

$$\mathbf{P}^c = \mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{W}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}$$

with $\mathbf{P}^c$ the projection onto the orthogonal complement of $\mathbf{x}$ using the inner product defined by the weights matrix $\mathbf{W}$ for the model at convergence of the PIRLS algorithm. In practice, it may seem impractical for the definition of $\widetilde{\mathbf{B}}_{\text{sp}}$ to depend on the matrix $\mathbf{W}$ since the weights will only be known after the model is fitted (and the model matrix is needed to define the model). However, the fitted values in the spatial model and the orthogonalised model are expected to be similar as the model matrices have the same column space. Therefore, we may assume that the weights in the two models are also broadly the same so that, in practice, we can use the weights from the fitted spatial model in the definition of $\widetilde{\mathbf{B}}_{\text{sp}}$.

The generalised RSR model corresponds to a Gaussian model with model matrix $\check{\mathbf{X}} =$

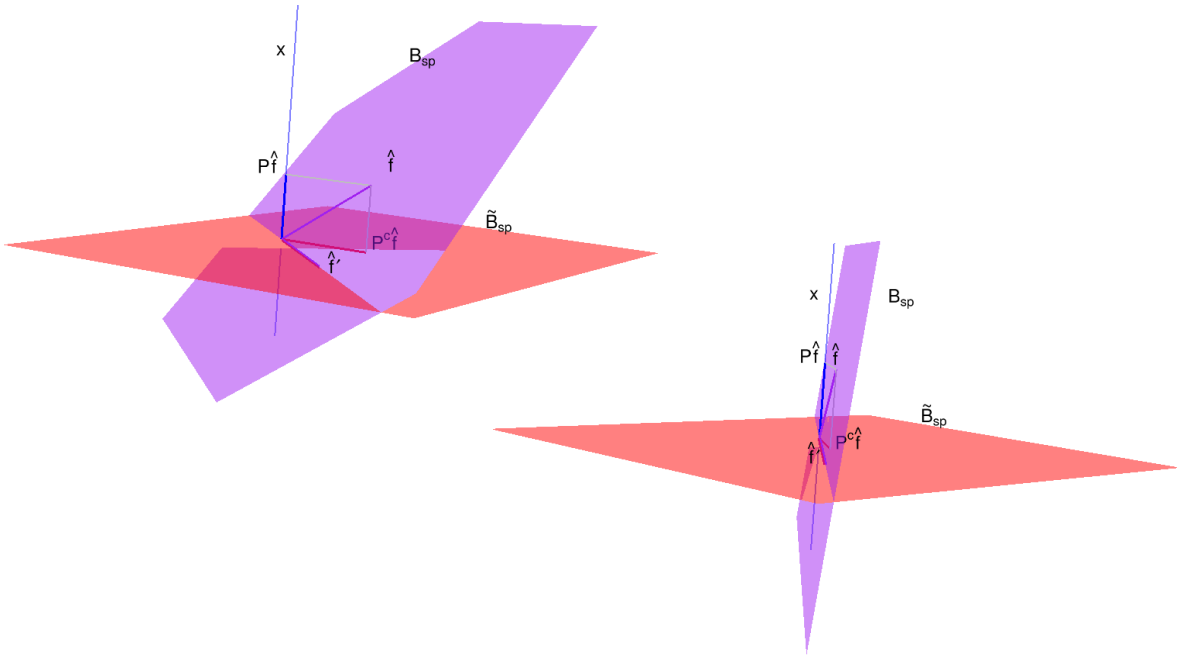


Figure 2-1: *For the two scenarios moderate collinearity (left) and high collinearity (right) between* $\mathbf{x}$ *and* $\mathbf{B}_{\text{sp}}$*, the figure shows the estimated spatial effect* $\widehat{\mathbf{f}}$ *in the spatial model decomposed as* $\widehat{\mathbf{f}} = \mathbf{P}\widehat{\mathbf{f}} + \mathbf{P}^c\widehat{\mathbf{f}}$ *with* $\mathbf{P}\widehat{\mathbf{f}}$ *in the span of* $\mathbf{x}$ *and* $\mathbf{P}^c\widehat{\mathbf{f}}$ *in* $\widetilde{\mathbf{B}}_{\text{sp}}$*. The component* $\mathbf{P}^c\widehat{\mathbf{f}}$ *does not lie in* $\mathbf{B}_{\text{sp}}$*, unless* $\mathbf{P}\widehat{\mathbf{f}} = \mathbf{0}$ *and* $\mathbf{P}^c\widehat{\mathbf{f}} = \widehat{\mathbf{f}}$*, i.e. unless* $\widehat{\mathbf{f}}$ *is itself orthogonal to the covariate (in which case,* $\widehat{\mathbf{f}}$ *would be like* $\widehat{\mathbf{f}}'$ *in the figure). We see that* $\mathbf{P}^c\widehat{\mathbf{f}}$ *can be quite different from* $\widehat{\mathbf{f}}$*, especially when* $\mathbf{x}$ *and* $\mathbf{B}_{\text{sp}}$ *are highly collinear.*

$[\check{\mathbf{x}}|\check{\widetilde{\mathbf{B}}}_{\text{sp}}]$ whose columns $\check{\mathbf{x}} = \sqrt{\mathbf{W}}\mathbf{x}$ and $\check{\widetilde{\mathbf{B}}}_{\text{sp}} = \sqrt{\mathbf{W}}\widetilde{\mathbf{B}}_{\text{sp}}$ are orthogonal by construction as

$$\check{\mathbf{x}}^T\check{\widetilde{\mathbf{B}}}_{\text{sp}} = \mathbf{x}^T\mathbf{W}\widetilde{\mathbf{B}}_{\text{sp}} = \mathbf{0}.$$

In line with (2.17), the estimated covariate effect $\hat{\check{\beta}}$ in this model is given by

$$\hat{\check{\beta}} = (\check{\mathbf{x}}^T\check{\mathbf{x}})^{-1}\check{\mathbf{x}}^T\check{\mathbf{y}} = (\mathbf{x}^T\mathbf{W}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}\mathbf{z}$$

This expression is exactly the same as the estimate of β that we would obtain in the corresponding null model, i.e. the GLM defined by $g(\mu_i) = \beta x_i$, $i = 1, \ldots, n$. However, we note that, although the expressions agree, the value of the estimate may differ as the fitted values, and therefore the weights $\mathbf{W}$ and pseudodata $\mathbf{z}$, in the two models do not generally agree. The fitted values in the RSR model are expected to be similar to those of the spatial model which has higher explanatory power than the null model and, therefore, is likely to have fitted values closer to the data. In the case where the true unmeasured spatial effects are large, the difference in fitted values could be significant as these effects would be explained in the spatial and RSR models, whereas in the null model they may be treated as residual noise. Hence, if the purpose of orthogonalisation is to preserve the null estimate, this would not necessarily be achieved in the non-Gaussian case.

Non-linear covariate effects

Spatial regression models often assume that covariate effects are linear, however, this is not always sufficient for modelling complex dependencies between the covariates and the response. In this context, another advantage of the thin plate spline formulation (2.2) of the spatial model is that it fits into the generalized additive models (GAMs) framework. GAMs, which were first developed by Hastie and Tibshirani (see Hastie and Tibshirani [1990]), are a very flexible non-parametric extension of linear models and GLMs that allow us to estimate unknown and possibly non-linear relationships between response and predictors. Just like the spatial effect in the formulation (2.2) is an unknown smooth function of spatial location, for each covariate in a GAM, the effect on the response variable is represented by an unknown smooth function defined on the domain of the covariate. That is, if $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\mathbf{x} = (x_1, \ldots, x_n)^T$ are the observed response and covariate as before, measured at spatial locations $\mathbf{t}_1, \ldots, \mathbf{t}_n$, the model is given by

$$y_i = f_{\text{co}}(x_i) + f_{\text{sp}}(\mathbf{t}_i) + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2). \tag{2.18}$$

where the unknown functions f_{co} and f_{sp} are estimated through penalised maximum likelihood with a separate smoothing penalty applied to each function.

More specifically, if $x_1, \ldots, x_n$ lie in the domain Ω_{co}, let $\mathbf{B}_{\text{co}} = [\mathbf{b}_1|\cdots|\mathbf{b}_k]$ denote a matrix whose columns are of the form $\mathbf{b}_j = (b_j(x_1), \ldots, b_j(x_n))^T$ where b_j denotes the j'th basis function for the space of functions defined on Ω_{co}. Thus, any function of the covariate f_{co} evaluated at the observed values of the covariate can be written as $\mathbf{f}_{\text{co}} = \mathbf{B}_{\text{co}}\boldsymbol{\beta}_{\text{co}}$. The GAM version of the spatial model (2.6) is then given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{2.19}$$

with model matrix $\mathbf{X} = [\mathbf{B}_{\text{co}}|\mathbf{B}_{\text{sp}}]$ and where the unknown coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\text{co}}^T, \boldsymbol{\beta}_{\text{sp}}^T)^T$ are estimated as the minimisers of

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta}$$

with penalty matrix

$$\mathbf{S} = \begin{bmatrix} \lambda_{\text{co}}\mathbf{S}_{\text{co}} & \mathbf{0} \\ \mathbf{0} & \lambda_{\text{sp}}\mathbf{S}_{\text{sp}} \end{bmatrix}$$

where $\lambda_{\text{co}} > 0$ and $\lambda_{\text{sp}} > 0$ are smoothing parameters estimated from a separate criterion as before. The normal equations then lead to the estimates

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_{\text{co}} \\ \widehat{\boldsymbol{\beta}}_{\text{sp}} \end{bmatrix} = (\mathbf{X}^T\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} \mathbf{B}_{\text{co}}^T\mathbf{B}_{\text{co}} + \lambda_{\text{co}}\mathbf{S}_{\text{co}} & \mathbf{B}_{\text{co}}^T\mathbf{B}_{\text{sp}} \\ \mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{co}} & \mathbf{B}_{\text{sp}}^T\mathbf{B}_{\text{sp}} + \lambda_{\text{sp}}\mathbf{S}_{\text{sp}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{B}_{\text{co}}^T \\ \mathbf{B}_{\text{sp}}^T \end{bmatrix} \mathbf{y}.$$

From this we obtain the covariate and spatial effect estimates $\widehat{\mathbf{f}}_{\text{co}} = \mathbf{B}_{\text{co}}\widehat{\boldsymbol{\beta}}_{\text{co}}$ and $\widehat{\mathbf{f}}_{\text{sp}} = \mathbf{B}_{\text{sp}}\widehat{\boldsymbol{\beta}}_{\text{sp}}$.

Note that there is some flexibility in model specification here as different choices of bases $\mathbf{B}_{\text{co}}$ and penalty structures $\mathbf{S}_{\text{co}}$ lead to different types of smooth structures for the covariate effect. While we have formulated the model for a one-dimensional covariate domain Ω_{co}, this domain could also be of higher dimension (as it is for the spatial effect). Moreover, in a similar way to the linear effects case, the model generalises to response distributions from the exponential family of distributions. A detailed account of GAM theory including different choices of smoothers can be found in Wood [2017].

Spatial confounding in the context of models of the form (2.18) becomes more complex. Having replaced the single column $\mathbf{x}$ in the model matrix by $\mathbf{B}_{\text{co}}$, which typically has many columns, collinearity between the covariate and spatial parts of the model may be more likely to arise. And estimates will be affected by smoothing, not only of the spatial part of the model, but also that of the covariate. In order to illustrate how spatial confounding presents itself when covariate effects are non-linear, we have simulated data in a similar way to our simulation study in Paper 1.

Using the same observed spatial fields $\mathbf{z} = (z_1, \ldots, z_n)^T$ and $\mathbf{z}' = (z'_1, \ldots, z'_n)^T$ as in the paper, we generate covariate data $\mathbf{x}$ and response data $\mathbf{y}$ as follows

$$\begin{aligned} \mathbf{x} &= 0.5\mathbf{z} + \boldsymbol{\epsilon}^x, \quad \boldsymbol{\epsilon}^x \underset{\text{iid}}{\sim} N(\mathbf{0}, \sigma_x^2) \\ \mathbf{y} &= \mathbf{f}_{\text{co}} + \mathbf{f}_{\text{sp}} + \boldsymbol{\epsilon}^y, \quad \boldsymbol{\epsilon}^y \underset{\text{iid}}{\sim} N(\mathbf{0}, \sigma_y^2) \end{aligned}$$

where $\sigma_x = 0.1$ and $\sigma_y = 1$ and true covariate effect given by $\mathbf{f}_{\text{co}} = (f_{\text{co}}(x_1), \ldots, f_{\text{co}}(x_n))^T$ with $f_{\text{co}}(x) = 3x^3$. We simulate this data for two different values of the true unmeasured spatial effect, namely, $\mathbf{f}_{\text{sp}} = -3\mathbf{z} - 0.5\mathbf{z}'$ (corresponding to high correlation with the covariate) and $\mathbf{f}_{\text{sp}} = -0.5\mathbf{z} - 3\mathbf{z}'$ (corresponding to low correlation).

Figure 2-2 shows the resulting estimated covariate smooth $\hat{f}_{\text{co}}$ in the null model (i.e. the model (2.18) with no spatial term) and the spatial model (2.18) for 100 independent runs of the simulation, using basis sizes $k_{\text{co}} = 10$ for the covariate smooth and $k_{\text{sp}} = 300$ for the spatial effect. We have fitted the spatial model with and without the smoothing penalty applied to the spatial effect. In the null model, the estimated covariate smooth is clearly biased when correlation between the covariate and the spatial effect is high but, when correlation is low, the model broadly captures the true effect, although with relatively high uncertainty. Without spatial smoothing, the spatial model estimate looks broadly unbiased, irrespective of the correlation between the covariate and the spatial effects. However, in the

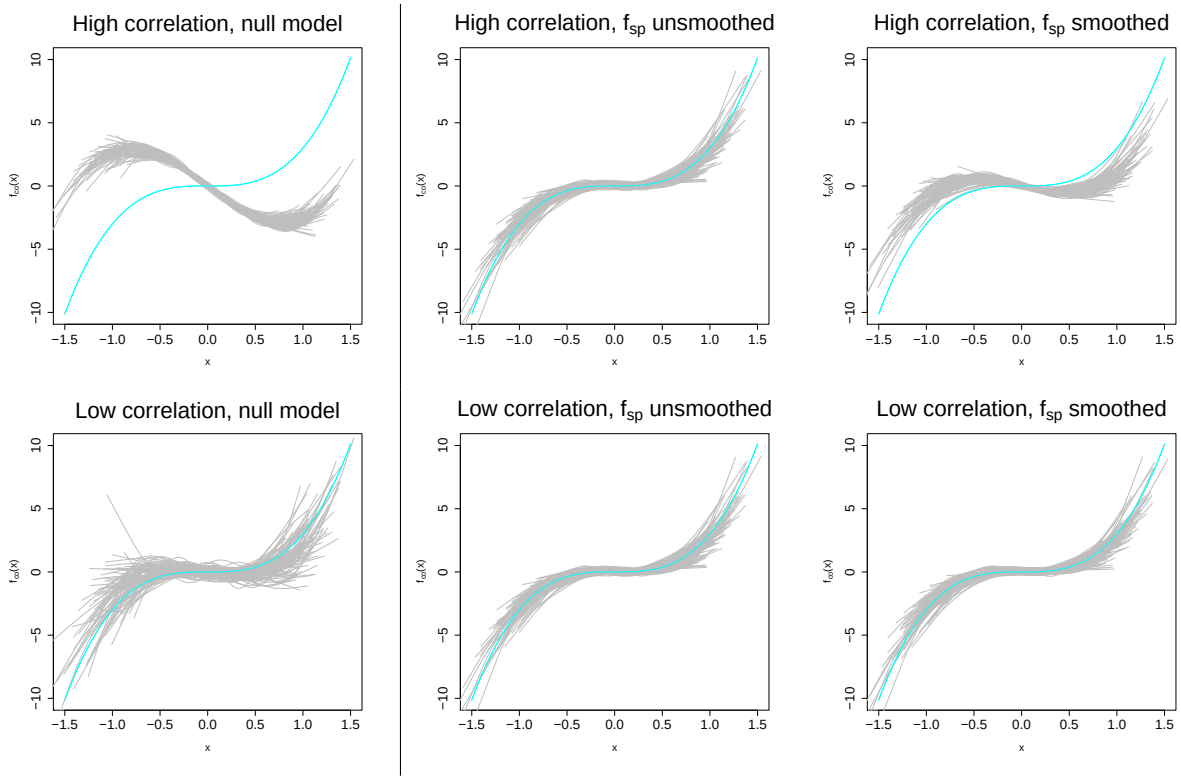


Figure 2-2: *Estimated covariate smooth $\hat{f}_{co}$ in the null model (left) and spatial model (middle and right) fitted to 100 data replicates, where the true covariate effect is $f_{co}(x) = 3x^3$ (shown in cyan). The spatial model has been fitted with (right) and without (middle) a smoothing penalty applied to the spatial effect. Results are shown in the case where correlation between the covariate and the true unmeasured spatial effect is high (top) and low (bottom).*

smoothed version of the model, the estimated smooth is no longer able to capture the true covariate effect when correlation is high. Thus, the overall behaviour is similar to what we saw in the linear effects case. The null model is biased as the covariate term will reflect, not only the effect of the covariate, but also any unmeasured spatial effects that are correlated with the covariate. And bias in the spatial model arises due to the combined effect of collinearity and spatial smoothing.

RSR in the non-linear setting can be implemented, once again, by replacing $\mathbf{B}_{\text{sp}}$ in the model matrix by the projected basis $\widetilde{\mathbf{B}}_{\text{sp}}$ but where $\mathbf{x}$ is now replaced by $\mathbf{B}_{\text{co}}$ in the definition of the projection $\mathbf{P}^c$. This recovers the effect estimate in the null model (at least when the response distribution is Gaussian). Of course, the resulting estimated covariate effect in the RSR model suffers from the same issues as those identified in the linear effects case, in particular the method introduces rather than removes bias when covariate and spatial effects are correlated.

It would be natural to try to extend the spatial+ method to the non-linear effects case. Using the same spatial regression as before to identify the spatial pattern $\widehat{\mathbf{f}}^x$ of the covariate $\mathbf{x}$, we once again obtain the decomposition $\mathbf{x} = \widehat{\mathbf{f}}^x + \mathbf{r}^x$ with $\widehat{\mathbf{f}}^x$ in $\mathbf{B}_{\text{sp}}$ and $\mathbf{r}^x$ the residuals. It seems tempting to simply replace $\mathbf{x}$ by $\mathbf{r}^x$ in the model (2.18) as we did before. However, the problem is that the residuals $\mathbf{r}^x$ no longer capture the covariate effect. To see this, suppose we have scalars y, x, s and r such that $x = s + r$ and the relationship betwen x and y is linear, say, $y = \beta x$. Then since $y = \beta(s + r) = \beta s + \beta r$, the effect β can be recovered from the relationship between r and y. But if $y = f(x) = f(s + r)$ for some non-linear function f,

unless $f(s+r) = f(s) + f(r)$, we cannot conclude that $y = f(s) + f(r)$ and the relationship between y and r may therefore be quite different to that described by f. Therefore, replacing $\mathbf{x}$ by the spatial residuals $\mathbf{r}^x$ in the model (2.18) and estimating the effect of $\mathbf{r}^x$ on $\mathbf{y}$ does not necessarily tell us much about the function f_{co}.

Conclusions

Using a thin plate spline formulation of the spatial model, we have studied the problem of spatial confounding in the case where covariates of interest are spatially dependent but not fully determined by spatial location. We see that the bias in the covariate effect estimate, in this case, arises as a direct result of spatial smoothing as, without smoothing, effect estimates in the model are all unbiased. This may seem surprising since work on spatial confounding tends to focus on collinearity issues and, as the smoothing penalty is only applied to the spatial part of the model, one might expect the covariate effect estimate to be largely unaffected by smoothing. However, as our analysis shows, unless the covariate is independent of the spatial basis vectors, then its effect estimate depends on the estimated spatial effect which, in turn, changes with smoothing.

In Paper 1 we propose a novel method, spatial+, for dealing with spatial confounding in this context. Detailed proofs of the theoretical results backing the method are provided in the paper, however, in this chapter we have shown that the intuition behind the model can be gained through simple linear algebra. Spatial+ is motivated by two observations. Firstly, we can remove the spatial part of the covariate in the spatial model and still identify the covariate effect and, secondly, in doing so, we decouple the covariate effect estimate from the spatial part of the model, making it much less sensitive to smoothing. Thus by replacing the covariates in the spatial model by their residuals after spatial dependence is regressed away, we obtain a model in which the estimate of the covariate effect stays largely unbiased when the spatial effect is smoothed.

For the RSR model, our conclusions agree with recent results in the literature [Khan and Calder, 2020, Nobre et al., 2020] that show that this method, in fact, creates rather than removes confounding bias. We see that by constraining the spatial effects, you essentially force each covariate term in the model to represent, not only the covariate of interest, but any unmeasured effects with a similar spatial pattern. As a result, the estimate will be biased by construction unless the unmeasured spatial effects are independent of the covariate. It may then seem appropriate to use RSR in the special case where we expect the covariates in the model to be independent of the true unmeasured spatial effects (as the resulting null effects would then be unbiased). However, in this case, smoothing in the spatial model also does not affect the covariate effect estimates and, therefore, there would be no need for any adjustments to the spatial model. Another problem with RSR is that the estimated orthogonalised spatial effect cannot, in general, be interpreted as "spatial" as it no longer lies in the space spanned by the spatial basis vectors. Finally, while it is possible to implement RSR for non-Gaussian response distributions, if the purpose of the orthogonalisation is to preserve the covariate effect estimate in the null model, then this is not always achieved. This is because, the estimates depend on weights and pseudodata of the fitted model that, in turn, depend on the fitted values which could be quite different in the null and RSR models.

Our investigation into spatial models implemented as GAMs with non-linear covariate effects shows that the problem of spatial confounding persists in these models, and it is once again the result of the combined effect of collinearity and spatial smoothing. However, the

analysis becomes more complicated as, in addition to spatial smoothing, a separate smoothing penalty is also applied to the covariate, leading to a more complex interaction between the covariate and spatial parts of the model. The spatial+ approach does not transfer directly to the non-linear setting as the ability to identify the covariate effect from its spatial residuals relies on the linearity of the effect. Thus, more work is needed to develop methodology that may reduce bias in effect estimates in spatial models of this type.

Chapter 3

Paper 1 - Spatial+: a novel approach to spatial confounding

Introduction to Paper 1

In this paper we consider spatial models with linear covariate effects for covariates that decompose into a smooth spatial part plus a residual. The spatial random effects are implemented as a thin plate spline as described in Chapter 2. In Chapter 2 we used linear algebra to analyse the behaviour of covariate effect estimates in these models and showed that bias arises due to the combination of collinearity and spatial smoothing. Here, we use asymptotic analysis and simulations to study this behaviour in more detail and propose a novel method, spatial+, for avoiding the bias in practice.

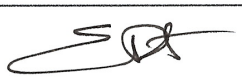
The thin plate spline formulation of the spatial model means that we can view it as a higher-dimensional version of a so-called partial spline model, i.e. a semiparametric model for which the domain of the spline (in our context, the spatial domain) is one-dimensional. For the one-dimensional case, Utreras [1983] studied the asymptotic properties of the smoother matrix $\mathbf{A}_\lambda$ in the expression (2.2), and Rice [1986] used these results to show that the bias in estimated covariate effects can become disproportionately large when the level of smoothing is chosen to optimise the model fit. Moreover, Chen and Shiau [1991] showed that this bias can be avoided by using an alternative two-stage spline smoothing model to estimate the covariate effect.

The understanding that spatial confounding for one-dimensional models in this setting is the same smoothing-induced bias as that discovered by Rice [1986] is key to this paper. Using results by Utreras [1988] on the asymptotic behaviour of thin plate splines in dimensions $d \geq 1$, we are able to generalise the results of Rice, Chen and Shiau to models of arbitrary spatial dimension. More specifically, we define a higher-dimensional analogue, spatial+, of the two-stage spline smoothing model and show that, for all spatial dimensions $d \geq 1$, the smoothing-induced bias in the spatial model persists while the bias in the spatial+ estimate is negligible. Moreover, we show that this approach generalises to models for which the response distribution is from the exponential family of distributions.

The paper includes a simulation study which illustrates the theoretical results for both Gaussian and non-Gaussian response distributions. We also apply the spatial+ method to forestry data.

Paper 1

Statement of Authorship

This declaration concerns the article entitled:			
Spatial+: a novel approach to spatial confounding			
Publication status (tick one) **Draft manuscript** [ ] **Submitted** [ ] **In review** [x] **Accepted** [ ] **Published** [ ]			
Publication details (reference)	Biometrics		
Copyright status (tick the appropriate statement) I hold the copyright for this material [x] Copyright is retained by the publisher, but I have been given permission to replicate the material here [ ]			
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	The work for this article was executed by the candidate under guidance from the co-authors. The candidate was the primary contributor to: Formulation of ideas and design/implementation of methodology: - The candidate had the main underlying idea for the methodology (to modify the spatial model, replacing covariates by spatial residuals) 100% - Presented with the idea that theoretical results from one-dimensional models may generalise to higher dimensions, the candidate independently formulated these generalisations and implemented all their proofs. 90% - Presented with the idea that the model may generalise to non-Gaussian distributions, the candidate independently formulated this generalisation. 90% - The candidate predominantly designed and implemented the simulation study and the data example. 90% Presentation of data in journal format: - The candidate wrote up the article for journal submission. The version of the article presented here is a revision following the comments from an associate editor and two referees. This version of the article was also written by the candidate. 100%		
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.		
Signed		**Date**	13/2/2021

Biometrics 64, 1–23 DOI: 10.1111/j.1541-0420.2005.00454.x
July 2020

Spatial+: a novel approach to spatial confounding

Emiko Dupont[1,*], Simon N. Wood[2], and Nicole H. Augustin[2]

[1]Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, U.K

[2]School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, U.K.

**email:* e.dupont@bath.ac.uk

Summary: In spatial regression models, collinearity between covariates and spatial effects can lead to significant bias in effect estimates. This problem, known as spatial confounding, is encountered modeling forestry data to assess the effect of temperature on tree health. Reliable inference is difficult as results depend on whether or not spatial effects are included in the model. We propose a novel approach, spatial+, for dealing with spatial confounding when the covariate of interest is spatially dependent but not fully determined by spatial location. Using a thin plate spline model formulation we see that, in this case, the bias in covariate effect estimates is a direct result of spatial smoothing. Spatial+ reduces the sensitivity of the estimates to smoothing by replacing the covariates by their residuals after spatial dependence has been regressed away. Through asymptotic analysis we show that spatial+ avoids the bias problems of the spatial model. This is also demonstrated in a simulation study. Spatial+ is straightforward to implement using existing software and, as the response variable is the same as that of the spatial model, standard model selection criteria can be used for comparisons. A major advantage of the method is also that it extends to models with non-Gaussian response distributions. Finally, while our results are derived in a thin plate spline setting, the spatial+ methodology transfers easily to other spatial model formulations.

Key words: Bias reduction; Collinearity; Forestry; Partial thin plate spline regression; Spatial confounding.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Regression models for spatially referenced data use spatial random effects to account for residual spatial correlation in the response data. As first noted by Clayton et al. (1993), these models can be problematic when estimation of individual covariate effects are of interest. So-called spatial confounding arises because spatial random effects may be correlated with spatially dependent covariates in the model and therefore interfere with their effect estimates. Reich et al. (2006) analyzed the issue using an example modeling the effect of socio-economic status on stomach cancer incidence in the municipalities of Slovenia. When spatial effects are added to the model, the covariate effect disappears, suggesting the spatial effects have taken over a disproportionate part of the explanatory power. While in this example, the spatial effects take the form of an Intrinsic Conditional Auto-Regressive (ICAR) random effect, spatial confounding is widely acknowledged as an issue that affects spatial models in general (see e.g. Hodges and Reich, 2010; Paciorek, 2010).

In this paper we model data from the Terrestrial Crown Condition Inventory (TCCI) forest health monitoring survey which has been carried out yearly since 1983 by the Forest Research Institute Baden-Württemberg. Crown defoliation (an indicator of poor tree health) has generally been worsening over time, and there is growing interest in understanding the effects of climate change in order to decide on forest management strategies for mitigation. Here, using a linear regression model, we consider the effect of temperature on crown defoliation. However, our results are highly dependent on whether or not we include spatial random effects in the model. As illustrated in Figure 1, in the null model (with no spatial effects), the estimated covariate effect is positive but not significant, whereas in the corresponding spatial model, the covariate effect is significant and the effect size more than triples. This behaviour suggests there is spatial confounding and makes reliable inference difficult.

[Figure 1 about here.]

A commonly used method for dealing with spatial confounding is restricted spatial regression (RSR), introduced by Reich et al. (2006) for the ICAR model, and further developed by Hanks et al. (2015) for continuous space models. In RSR the spatial random effects are restricted to the orthogonal complement of the covariates while keeping the overall column space of the model matrix in the regression unchanged. RSR directly eliminates collinearity and is designed to preserve the estimate of the null model while still accounting for residual spatial correlation. However, in the presence of unmeasured spatial confounders, the RSR estimate of the covariate effect may be significantly biased as it reflects not only the effect of the covariate but also that of the confounders (see e.g. Hanks et al., 2015; Khan and Calder, 2020). Here, we define a spatial confounder in the classical sense of an unmeasured spatial variable (causal or otherwise) that is associated with both the covariate and the response (Kirkwood and Sterne, 2010; McNamee, 2003).

Paciorek (2010) and Page et al. (2017) study the behaviour of the estimates in the spatial model when the covariate, like the response variable, has a spatial covariance structure. Intuitively, the model cannot distinguish the covariate from an unmeasured spatial effect, and the apportionment of effects between the covariate and spatial parts of the model may therefore be somewhat arbitrary. The analysis shows that the size of the resulting bias in the covariate effect estimate depends on the relative spatial scales of the covariate and spatial effects and, when the spatial scales agree, the bias is the same as that of RSR. Thus, while the estimate in the spatial model differs from RSR, it may be just as biased.

In many practical applications, however, the covariate of interest is spatially dependent but not fully determined by spatial location. This form of the covariate is assumed in Thaden and Kneib (2018) who propose the geoadditive structural equations model (gSEM). Here, spatial dependence is regressed away from both the response and the covariates, and a regression involving the residuals only is used to identify the original covariate effect. Simulations show

that the bias in the covariate effect estimate of the spatial model is broadly removed using the gSEM. However, it is not immediately clear why the method works, and when the variables of interest are naturally spatially dependent, it seems undesirable to eliminate all spatial information from the modeling. The change in response variable also means that standard model selection criteria cannot be used for comparisons with the spatial and null models.

The structure of the covariate is usually not highlighted in the spatial confounding literature, but is important, as non-spatial information in a covariate can be used to distinguish it from the spatial effects without the need for considering differences in spatial scales. In this paper we show that, when a covariate is not fully spatial, unmeasured spatial confounders may still lead to significant bias in its effect estimate in the spatial model, however, the bias can be avoided in a relatively straightforward way. We propose a novel approach, spatial+, that is a simple modification of the spatial model in which the covariate is replaced by its residuals after spatial dependence has been regressed away. Similar to RSR, spatial+ retains the column space of the model matrix while reducing collinearity, but by adjusting the covariate rather than the spatial part of the model. Using asymptotic analysis as well as a simulation study we show that the estimates in spatial+ avoid the bias problems of the spatial model. We note that our asymptotic analysis applied to the gSEM estimates confirm the results of Thaden and Kneib (2018) and, for completeness, these derivations are included in Web Appendix D. An advantage of spatial+, however, is that all spatial information is retained in the model. Moreover, while the main properties of spatial+ are studied for models with a Gaussian response variable, we show that the method generalizes naturally to any response distribution from the exponential family of distributions.

Key to our analysis is that we formulate the spatial model as a partial thin plate spline model. Here, spatial correlation is modeled by imposing a smoothing penalty on the spatial effects in the fitting process. We then see that the bias in the covariate effect estimate

arises as a direct result of smoothing, and spatial+ is a modification of the model matrix that makes the covariate part of the model less sensitive to this. Although our results are derived in the thin plate spline context, the methodology of spatial+ can be directly applied to other commonly used spatial models including, for example, Gaussian Markov random field (GMRF) models and the (discrete space) ICAR model. In fact, it can be shown that modeling spatial random effects through the use of a smoothing penalty is equivalent to a Bayesian model formulation in which the spatial correlation structure is determined by a prior distribution. This equivalence is explained, for example, in Kimeldorf and Wahba (1970), Section 6.1 of Silverman (1985), pages 239-240 of Wood (2017) and Fahrmeir et al. (2004). Thus, while different spatial models correspond to different smoothing penalties, the underlying idea of reducing collinearity in this way to keep covariate effect estimates unaffected by spatial smoothing would apply in general.

This paper is structured as follows. In Section 2, we introduce the spatial and spatial+ models that form the basis of our analysis. Section 3 summarizes our asymptotic analysis, details of which are in the supplementary web material. In Section 4, we illustrate our theoretical results in a simulation study which also compares spatial+ with RSR and the gSEM. In Section 5, we demonstrate how spatial+ can be implemented by applying it to our forestry example. Finally, in Section 6, we generalize the spatial+ methodology to non-Gaussian response distributions and confirm that the method works in simulations for three different distributions.

2. Method

2.1 *Spatial model*

Our starting point is a spatial model formulated as a partial thin plate spline model (see e.g. Wahba, 1990) of the form

$$y_i = \beta x_i + f(\mathbf{t}_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the response, $\mathbf{x} = (x_1, \ldots, x_n)^T$ an observed covariate with unknown effect β and f an unknown bounded function defined on an open bounded domain $\Omega \subset \mathbb{R}^d$ which includes the data locations $\mathbf{t}_1, \ldots, \mathbf{t}_n$. The estimates $\widehat{\beta}$ and $\widehat{f}$ in this model (known as the partial thin plate spline estimates of order $m > d/2$) are obtained as the minimizers of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta x_i - f(\mathbf{t}_i))^2 + \lambda \sum_{i_1,\ldots,i_m} \int_{\mathbb{R}^d} \Big|\frac{\partial^m f(\mathbf{t})}{\partial t_{i_1} \cdots \partial t_{i_m}}\Big|^2 d\mathbf{t}$$

where $\lambda > 0$ is an unknown smoothing parameter. Minimization here is over all $\beta \in \mathbb{R}$ and functions $f \in H^m(\mathbb{R}^d)$ with $\frac{\partial^m f}{\partial t_{i_1} \cdots \partial t_{i_m}} \in L^2(\mathbb{R}^d)$ for all subsets $i_1, \ldots, i_m$ of $1, \ldots, n$. The first term encourages fitted values that are close to the data while the second term induces smoothing by penalizing the wiggliness of the function f.

Duchon (1977) showed that the estimate of f can be obtained by estimating its coefficients in a basis known as the natural thin plate spline basis. This basis spans a finite-dimensional subspace in the space of functions defined on $\mathbb{R}^d$ and has dimension $N = M + n$ where $M = \binom{m+d-1}{d}$. Using this basis, the partial thin plate spline estimates $\widehat{\beta}$ and $\widehat{\mathbf{f}} = (\widehat{f}(\mathbf{t}_1), \ldots, \widehat{f}(\mathbf{t}_n))^T$ are the minimizers of

$$\|\mathbf{y} - \beta\mathbf{x} - \mathbf{f}\|^2 + n\lambda \mathbf{f}^T \mathbf{\Gamma} \mathbf{f} \tag{2}$$

with $\mathbf{\Gamma}$ an $n \times n$ diagonal penalty matrix. Solving the resulting normal equations, we see that

$$\widehat{\beta} = \left(\mathbf{x}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{x}\right)^{-1} \mathbf{x}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}, \quad \widehat{\mathbf{f}} = \mathbf{S}_\lambda(\mathbf{y} - \widehat{\beta}\mathbf{x}) \tag{3}$$

where $\mathbf{S}_\lambda = (\mathbf{I} + n\lambda\mathbf{\Gamma})^{-1}$ is known as the smoother matrix.

2.2 *Spatial+ model*

Starting with the model (1), in line with Rice (1986), we assume the covariate $\mathbf{x}$ has the form

$$x_i = f^x(\mathbf{t}_i) + \epsilon_i^x, \quad \epsilon_i^x \overset{\text{iid}}{\sim} N(0, \sigma_x^2) \tag{4}$$

where $f^x \in H^m(\Omega)$ is bounded. This means that $\mathbf{x}$ is correlated with the smooth term f through the component f^x. Extending the two-stage smoothing spline model defined in Chen and Shiau (1991) to models of dimension $d \geqslant 1$, we define the spatial+ model as follows. Let $\widehat{\mathbf{f}}^x = \mathbf{S}_{\lambda_x}\mathbf{x}$ and $\mathbf{r}^x = (\mathbf{I} - \mathbf{S}_{\lambda_x})\mathbf{x}$ be the fitted values and residuals in the thin plate spline regression (4) with smoothing parameter $\lambda_x > 0$. The spatial+ model is then the partial thin plate spline model

$$y_i = \beta r_i^x + f^+(\mathbf{t}_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{5}$$

where β is the unknown effect of $\mathbf{r}^x = (r_1^x, \ldots, r_n^x)^T$ and f^+ models the combined effect $f + \beta f^x$ in the original model (1). The spatial+ estimate $\widehat{\beta}^+$ of β is its partial thin plate spline estimate in this model, i.e.

$$\begin{aligned}\widehat{\beta}^+ &= \left(\mathbf{r}^{xT}(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{r}^x\right)^{-1}\mathbf{r}^{xT}(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y} \\ &= \left(\mathbf{x}^T(\mathbf{I} - \mathbf{S}_{\lambda_x})(\mathbf{I} - \mathbf{S}_\lambda)(\mathbf{I} - \mathbf{S}_{\lambda_x})\mathbf{x}\right)^{-1}\mathbf{x}^T(\mathbf{I} - \mathbf{S}_{\lambda_x})(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y},\end{aligned}$$

and the spatial+ estimate $\widehat{\mathbf{f}}^+$ of $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$ is given by

$$\widehat{\mathbf{f}}^+ = \mathbf{S}_\lambda(\mathbf{y} - \widehat{\beta}^+\mathbf{x}) - \widehat{\beta}^+\widehat{\mathbf{f}}^x = \mathbf{S}_\lambda(\mathbf{y} - \widehat{\beta}^+\mathbf{x}) - (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\widehat{\beta}^+\mathbf{x}.$$

2.3 *Smoothness selection*

Smoothing penalties introduce bias in estimates but reduce variance. The smoothing parameters λ and λ_x are usually estimated based on a separate smoothness selection criterion that balances this bias-variance trade-off.

For the analysis summarized in Section 3, in line with Rice (1986); Chen and Shiau (1991), we choose the value of the smoothing parameter that minimizes the average mean squared error (AMSE) of the estimated spatial effect. The AMSE for an estimated effect $\widehat{\mathbf{f}} = (\widehat{f}_1, \ldots, \widehat{f}_n)^T$ of $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$ is defined as

$$\text{AMSE}(\widehat{\mathbf{f}}) = \frac{1}{n}\sum_{i=1}^{n} \text{E}\left[(\widehat{f}_i - f(\mathbf{t}_i))^2\right] = B^2(f, \lambda) + V(f, \lambda)$$

where $B^2(f, \lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\text{E}(\widehat{f}_i) - f(\mathbf{t}_i)\right)^2$ and $V(f, \lambda) = \frac{1}{n}\sum_{i=1}^{n}\text{Var}(\widehat{f}_i)$ are the average

squared bias and the average variance, respectively. (Note that $\widehat{\mathbf{f}}$ is a linear transformation of $\mathbf{y}$ and therefore has a multivariate normal distribution).

For the simulations in Section 4 and the example in Section 5 we use the generalized cross validation (GCV) criterion, which is the default option in the `R`-package `mgcv` used for implementation. Asymptotically (as the sample size $n \to \infty$), GCV selects the optimal smoothing parameter for minimizing prediction error. Thus, GCV is not dissimilar to the criterion used for the theoretical derivations. Indeed, Chen and Shiau (1994) show that their asymptotic results in Chen and Shiau (1991) for one-dimensional models also hold for GCV and Mallows' C_L. In practice, the restricted maximum likelihood (REML) criterion is often used instead of GCV as, for finite samples, GCV usually has more uncertain estimates than REML and tends to undersmooth (i.e. overfit) the data (Wood (2017) p. 266-267). Repeating the simulations and the data example using REML gave similar results to GCV.

3. Asymptotic results

In the supplementary web material we analyze the asymptotic behaviour of effect estimates in the models defined in Sections 2.1 and 2.2 as the sample size $n \to \infty$. Without the smoothing penalty, the spatial model (1) is an ordinary linear model in which the estimates are unbiased. Therefore, bias in the covariate effect estimate $\widehat{\beta}$ arises as a direct result of smoothing. In fact, for partial spline models (i.e. models where the domain of the spline, here the spatial domain, is one-dimensional) Rice (1986) identified this smoothing-induced bias and showed that it can become disproportionately large unless the data is undersmoothed. More specifically, Rice's results show that if the smoothing parameter λ converges at the optimal rate (for minimizing the AMSE of the estimated smooth effect), it cannot be ensured that the bias of $\widehat{\beta}$ converges faster than its standard deviation. Spatial+ is a higher-dimensional version of a model introduced by Chen and Shiau (1991) to overcome this type of bias in dimension one.

Rice, Chen and Shiau use the Demmler-Reinsch basis for natural splines to diagonalize

the smoother matrix $\mathbf{S}_\lambda$, enabling them to explicitly study the asymptotic behaviour of the estimates in one-dimensional models. Due to results by Utreras (1988), we are able to extend these derivations to models of arbitrary spatial dimension. The main results of our analysis are provided in Web Appendix C. We confirm that, as is the case in dimension one, when smoothing is chosen at an optimal rate for minimizing the AMSE of the estimated spatial effect, the bias in the covariate effect estimate in the spatial model can become disproportionately large. In contrast, in the spatial+ model, the bias converges to 0 strictly faster than the standard deviation. Our results are based on a number of technical lemmas and assumptions, details of which are provided in Web Appendices A and B.

4. Simulation

Partial thin plate spline models can be implemented in the `R`-package `mgcv` using the computationally efficient reduced rank approximation known as thin plate regression splines. We use this implementation (with GCV as the smoothness selection criterion) to compare the results of models fitted to simulated data for which we know the true underlying covariate and spatial dependence.

4.1 *Data*

We generate 100 independent replicates of covariate data $\mathbf{x} = (x_1, \ldots, x_n)^T$ and response data $\mathbf{y} = (y_1, \ldots, y_n)^T$, observed at $n = 1000$ randomly selected locations in the spatial domain $[0, 10] \times [0, 10]$ in $\mathbb{R}^2$ (using a 50×50 grid), as follows. Let $\mathbf{z} = (z_1, \ldots, z_n)^T$ and $\mathbf{z}' = (z'_1, \ldots, z'_n)^T$ denote observations at the selected locations of independently generated Gaussian spatial fields with an exponential and a spherical covariance structure, respectively. That is, each spatial field is sampled from a multivariate normal distribution centered at $\mathbf{0}$ with covariance structure defined by $C(h) = \exp(-(h/R)^p)$ with $R = 5$ and $p = 1$ for the exponential field and $C(h) = -1 - 1.5h/R + 0.5(h/R)^3$ for $h \leqslant R$, $C(h) = 0$ for $h > R$ with $R = 1$ for the spherical field (where h denotes Euclidean distance). To ensure that the fields

lie in the span of the spatial basis vectors used for the models in Section 4.2, each is replaced by the fitted values of a spatial thin plate regression spline fitted to them. We then let

$$\begin{aligned} \mathbf{x} &= 0.5\mathbf{z} + \boldsymbol{\epsilon}^x \text{ where } \boldsymbol{\epsilon}^x \sim N(\mathbf{0}, \sigma_x^2 \mathbf{I}), \\ \mathbf{y} &= \beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}^y \text{ where } \boldsymbol{\epsilon}^y \sim N(\mathbf{0}, \sigma_y^2 \mathbf{I}), \end{aligned}$$

with true covariate effect $\beta = 3$, true residual spatial effect $\mathbf{f} = -\mathbf{z} - \mathbf{z}'$ and $\sigma_y = 1$, $\sigma_x = 0.1$. Thus, $\mathbf{f}$ is directly correlated with the spatial pattern $0.5\mathbf{z}$ of the covariate. This approach is similar to Thaden and Kneib (2018), except we have added the component $-\mathbf{z}'$ so that $\mathbf{f}$ could represent, for example, the combined effect of an unobserved covariate (with a similar spatial pattern to that of $\mathbf{x}$) as well as an independent short-range spatial process. Also, rather than treating the spatial fields as fixed, we generate new fields for each replicate in the simulation. This slight change in approach was chosen to show that results do not rely on any particular replicates of the spatial fields. We note that similar results were obtained when we repeated the simulations for a number of fixed spatial fields. Finally, we have chosen σ_x relatively small (such that the model matrix for the spatial model has nearly collinear columns) and σ_y relatively large (to encourage smoothing). This is the situation in which we would expect spatial confounding issues to arise which is also confirmed by the simulations in Thaden and Kneib (2018).

4.2 *Models*

To each replicate of simulated response data $\mathbf{y}$ and covariate data $\mathbf{x}$, we fit the following models (with basis size $k = 300$ for the thin plate regression splines). Models 2 - 5 are fitted twice: once with smoothing penalties applied (i.e. where λ and λ_x are estimated from the data) and once without (i.e. where $\lambda = \lambda_x = 0$). In `mgcv`, smoothing penalties are applied by default but can be removed using the option `fx=TRUE`. This option means that the smoothing parameter is fixed rather than estimated, defaulting to 0 (i.e. no smoothing) when no value is specified.

1. Null model: The model with no spatial effects given by

$$y_i = \beta x_i + \epsilon_i, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{6}$$

where β and σ^2 are estimated parameters.

2. Spatial model: The model given by

$$y_i = \beta x_i + f(\mathbf{t}_i) + \epsilon_i, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{7}$$

where β and σ^2 are estimated parameters and f a thin plate regression spline with $\mathbf{t}_1, \ldots, \mathbf{t}_n$ the observed data locations.

3. RSR model: Let $\mathbf{B}_{\text{sp}}$ be the matrix whose columns are the spatial basis vectors in the model matrix from (7) (i.e. the thin plate regression spline basis functions evaluated at the data locations) and let $\widetilde{\mathbf{B}}_{\text{sp}} = (\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{B}_{\text{sp}}$ be the projection of this onto the orthogonal complement of $\mathbf{x}$. The RSR model is given by

$$y_i = \beta x_i + \widetilde{f}_i + \epsilon_i, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{8}$$

where β and σ^2 are estimated parameters and $\widetilde{\mathbf{f}} = (\widetilde{f}_1, \ldots, \widetilde{f}_n)^T$ is modeled the same way as the spatial effect in (7) but with $\mathbf{B}_{\text{sp}}$ replaced by $\widetilde{\mathbf{B}}_{\text{sp}}$ in the model matrix.

4. gSEM: Let $\mathbf{r}^x = (r_1^x, \ldots, r_n^x)^T$ and $\mathbf{r}^y = (r_1^y, \ldots, r_n^y)^T$ denote the spatial residuals of $\mathbf{x}$ and $\mathbf{y}$, that is, $\mathbf{r}^x = \mathbf{x} - \widehat{\mathbf{f}}^x$ where $\widehat{\mathbf{f}}^x$ are the fitted values in the regression

$$x_i = f^x(\mathbf{t}_i) + \epsilon_i^x, \quad \boldsymbol{\epsilon}^x \sim N(\mathbf{0}, \sigma_x^2 \mathbf{I}) \tag{9}$$

where σ_x^2 is estimated and f^x a thin plate regression spline, and $\mathbf{r}^y$ is the same but replacing $\mathbf{x}$ by $\mathbf{y}$. The gSEM model is then the linear model given by

$$r_i^y = \beta r_i^x + \epsilon_i, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{10}$$

where β and σ^2 are estimated.

5. Spatial+: Let $\mathbf{r}^x$ denote the spatial residuals of $\mathbf{x}$ as above. The spatial+ model is then

$$y_i = \beta r_i^x + f^+(\mathbf{t}_i) + \epsilon_i, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{11}$$

where β and σ^2 are estimated parameters and f^+ a thin plate regression spline.

4.3 *Results*

The results of the simulation are summarized in Figure 2. For each data replicate, the output is the estimated covariate effect and the mean squared error (MSE) of fitted values for each model fit. For ease of notation, in this section, we use $\widehat{\beta}$ to mean the estimated covariate effect in any of the fitted models (rather than the partial thin plate spline estimate alone). The MSE of fitted values is calculated as $\|\widehat{\mathbf{y}} - (\beta\mathbf{x} + \mathbf{f})\|^2$ where for models 1, 2, 3 and 5, $\widehat{\mathbf{y}}$ is the fitted values in the regressions (6), (7), (8) and (11), respectively, and for model 4, $\widehat{\mathbf{y}} = \widehat{\mathbf{f}}^y + \widehat{\mathbf{r}}^y$ where $\widehat{\mathbf{f}}^y$ and $\widehat{\mathbf{r}}^y$ are the fitted values in the regressions (9) and (10). Here $\beta = 3$ and $\mathbf{f} = -\mathbf{z} - \mathbf{z}'$ are the true values of the estimated effects with $\beta\mathbf{x} + \mathbf{f}$ the true mean of $\mathbf{y}$.

[Figure 2 about here.]

In the null model and the RSR model, the estimated covariate effect is the same (that is, for any given data replicate, the value of $\widehat{\beta}$ is identical) and has a noticeably larger bias than the estimates in the other models. This is expected as for these models, $\widehat{\beta}$ is the ordinary least squares estimate, which, in addition to the true effect β, includes a contribution from the part of $\mathbf{f}$ that is correlated with $\mathbf{x}$. More specifically, the bias in $\widehat{\beta}$ is given by $\mathrm{E}((\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{f})$ (see Web Appendix E). Note that in our simulations, since $\mathbf{x}$ and $\mathbf{f}$ are negatively correlated, the bias is negative, however, if the correlation had been positive, the bias would have been positive. The fitted values in RSR, however, differ from those of the null model as the larger model matrix explains a part of $\mathbf{y}$ that is treated as random noise in the null model. In fact, the column space of the model matrix of the RSR model is the same as that of the spatial model, and it is therefore not surprising that, for any given data replicate, the fitted values in these two models are similar.

If no smoothing penalty is applied, models 2, 4 and 5 are essentially the same: for any given data replicate, they have the same fitted values and the same unbiased estimate for the covariate effect. This illustrates that spatial confounding bias is due to the combined

effect of collinearity and smoothing, rather than collinearity alone. The spatial model is, in this case, an ordinary linear model where the columns in the model matrix are the covariate $\mathbf{x}$ and the spatial basis vectors $\mathbf{B}_{\text{sp}}$. This is the model from which the data is generated and, therefore, it is not surprising that the spatial model is able to recapture the true effects. The spatial+ model is a reparametrization of the spatial model which preserves the overall column space, and simple linear model theory shows that the covariate effect estimate is preserved. Similarly, it is straightforward to show that the gSEM covariate effect estimate agrees with the spatial model in this case. (Derivations are included in Web Appendix E).

In the unsmoothed versions of models 2 - 5, for any given data replicate, the fitted values are the same across all models. When smoothing is applied, the MSE of fitted values reduces, indeed, this is the intended purpose of the smoothing penalty. Looking at the covariate effect estimate, in the RSR model, the (biased) ordinary least squares estimate is unaffected by smoothing. For the remaining three models, while the unsmoothed versions of the models give unbiased estimates of β, we see that smoothing introduces varying degrees of bias. In the spatial model, the bias is quite large illustrating our results from Section 3. In contrast, while the covariate effect estimate is no longer the same in the gSEM and the spatial+ model, for both models, the bias is still negligible. This behaviour is therefore also consistent with what we would expect from our theoretical results.

4.4 *Additional comments*

Our analysis in Section 4.3 gives some intuition for why spatial+ works. If no smoothing penalty is applied we saw that for any given data replicate, spatial+ has the same unbiased estimate for the covariate effect as the spatial model. In fact, any decomposition $\mathbf{x} = \mathbf{v} + \mathbf{r}$ with $\mathbf{v}$ in the column space of the spatial basis vectors $\mathbf{B}_{\text{sp}}$ gives a reparametrization (replacing $\mathbf{x}$ by $\mathbf{r}$) in which $\mathbf{r}$ captures the original covariate effect (since $\beta\mathbf{x} = \beta\mathbf{v} + \beta\mathbf{r}$). However, by choosing $\mathbf{r}$ to be broadly orthogonal to the column space of $\mathbf{B}_{\text{sp}}$ (as it is in the

spatial+ model), the estimates of the covariate and spatial effects are broadly decorrelated. Thus, the covariate effect estimate is largely unaffected when smoothing is applied to the spatial term and thereby remains broadly unbiased.

Although the asymptotic results in Section 3 are technically only expected to hold for large sample sizes n, the above intuition of spatial+ applies in general. In order to investigate the behaviour at moderate sample sizes, we repeated the simulations for $n = 300$, $n = 150$ and $n = 50$ (with spatial basis sizes $k_{\text{sp}} = 100$, $k_{\text{sp}} = 100$ and $k_{\text{sp}} = 30$, respectively). The results are included in Web Appendix F. We see that the smaller the sample size, the larger the variability of the estimate $\widehat{\beta}$, particularly for the unsmoothed spatial model, gSEM and spatial+ model as well as the smoothed gSEM and spatial+ model. However, overall, the behaviour of the bias looks broadly similar to the results for the larger sample size and are in line with the asymptotic results. We note that, when the spatial basis size k_{sp} is large compared to n, the MSE of fitted values in the unsmoothed models becomes relatively high. The thin plate regression spline basis is generally ordered with lower frequency spatial patterns first so that adding more spatial basis functions increases the ability of the spatial effect to model more complex spatial variation involving both lower and higher frequency spatial patterns. In practice, while a small basis size is preferable for reducing computation time for model fitting, a large basis may be necessary in order to capture the spatial variation in the data. However, when the flexibility of the spatial effect is increased, the model is also more likely to overfit the data. The purpose of smoothing is exactly to avoid this overfitting and reduce the effective degrees of freedom in the model.

Finally, our results assume that the true spatial dependence of both the response variable and the covariate can be described by thin plate splines. (In our simulations this was ensured by fitting thin plate splines to the spatial fields $\mathbf{z}$ and $\mathbf{z}'$ used for the data generation.) If this is not the case, there may be additional bias caused by model mis-specification. In practical

terms, the thin plate spline formulation assumes that spatial dependence is smooth and isotropic, but when these conditions hold it provides a fairly flexible way of modeling spatial effects. Thus, when we repeated our simulations using data based on spatial fields that were fitted with Gaussian process smooths rather than thin plate splines, the results were very similar. (See Web Appendix F).

5. Application

We illustrate how the spatial+ model can be used in practice by applying it to our forestry example. Details of the data can be found in Augustin et al. (2009); Eichhorn et al. (2017). We consider here the data for spruce for a single observation year, namely, 2013 which has measurements from $n = 186$ locations. We are interested in assessing the effect of the climate variable `tminmay` (minimum temperature in May) on the response variable `ratio` (crown defoliation expressed as a proportion). We expect a high minimum temperature in May to be indicative of a warmer and drier year in general which, in turn, is likely to lead to higher levels of tree defoliation (measured later in summer). We also expect older trees to have significantly more defoliation than younger trees and have therefore included the variable `age` (age of trees) as an additional covariate in the models. Scatterplots of the data (not shown here) indicate the relationships between the covariates and the response variable are broadly as expected.

5.1 *Models*

A natural starting point is the null model

$$\texttt{ratio}_i = \alpha + \beta_1 \texttt{age}_i + \beta_2 \texttt{tminmay}_i + \epsilon_i, \tag{12}$$

where $\epsilon_i \sim N(0, \sigma^2)$ is iid noise and α, β_1, β_2 and σ are estimated parameters. However, numerous spatially dependent predictors have not been included in the model, for example, soil characteristics such as soil depth and base saturation; other climatic variables such as

those related to radiation and precipitation; water budget of the trees etc. Therefore, we would expect residual spatial correlation in the response variable, and a more appropriate model may therefore be a spatial model, which we define as

$$\texttt{ratio}_i = \alpha + \beta_1 \texttt{age}_i + \beta_2 \texttt{tminmay}_i + f(\mathbf{t}_i) + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ is iid noise, α, β_1, β_2 and σ are estimated parameters and f a thin plate regression spline (with basis size $k = 100$) with $\mathbf{t}_1, \ldots, \mathbf{t}_n$ the observed data locations.

The covariate effects of interest are β_1 and β_2 but, as the results of Sections 3 and 4 show, the estimates of these effects may be highly biased in both the null model and the spatial model. This disproportionate bias is avoided in the spatial+ model. Let $\mathbf{r}^1 = (r_1^1, \ldots, r_n^1)^T$ and $\mathbf{r}^2 = (r_1^2, \ldots, r_n^2)^T$ be the residuals when a thin plate regression spline (with basis size $k = 100$) is fitted to `age` and `tminmay`, respectively. The spatial+ model is then

$$\texttt{ratio}_i = \alpha + \beta_1 r_i^1 + \beta_2 r_i^2 + f^+(\mathbf{t}_i) + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ is iid noise, α, β_1, β_2 and σ are estimated parameters and f^+ a thin plate regression spline (with basis size $k = 100$) with $\mathbf{t}_1, \ldots, \mathbf{t}_n$ the observed data locations.

Finally, for comparison, we fit the gSEM as an alternative method for avoiding spatial confounding bias. Let $\mathbf{r}^y = (r_1^y, \ldots, r_n^y)^T$ be the residuals when a thin plate regression spline (with basis size $k = 100$) is fitted to the response variable `ratio`. The gSEM is then

$$r_i^y = \beta_1 r_i^1 + \beta_2 r_i^2 + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ is iid noise and β_1, β_2 and σ are estimated parameters.

5.2 *Results*

The results of fitting the above four models to the data are summarized in Table 1.

[Table 1 about here.]

The spatial term in the spatial model is significant, which confirms there is residual spatial correlation in the data as expected. Furthermore, as the spatial term allows for more of

the residual variation to be explained, the deviance explained is higher and the estimated standard deviation is lower than in the null model. As the AIC is also lower, we conclude that the spatial model is an overall better fitting model than the null model for this data. However, while the spatial model may be appropriate for overall predictions of the response variable, the estimate of any individual covariate effect may be biased. Using the spatial+ model, we expect to obtain similar fitted values as the spatial model but with covariate effect estimates that have only negligible bias. Indeed, in terms of overall fit, we see that the deviance explained, estimated standard deviation and AIC in the spatial+ model are similar to those of the spatial model. For completeness, we have also included the gSEM. Note, however, that in the gSEM, since the response variable in the regression differs from that of the other three models, the deviance explained, estimated standard deviation and AIC cannot be directly compared to the other models.

The covariate `age` is highly significant and has a positive effect as expected. This covariate does not appear to be affected by spatial confounding as the estimated effect and its p-value are largely robust to the choice of model. This happens, for example, if a covariate is independent of the true underlying residual spatial effect. Also, in the case of `age`, not only is this a covariate that is not very well explained by spatial location (a spatial smooth fitted to this variable has deviance explained of only 13%), but its estimated spatial pattern looks dominated by linear spatial basis functions which are unpenalized in the spatial model. Therefore, penalization of the spatial term f in the spatial model is less likely to interfere with the covariate effect estimate (see Rice (1986) Proposition D).

In contrast, the estimated effect of the covariate `tminmay` is not significant in the null model but is significant in the spatial model and is even more significant in the spatial+ model. Furthermore, while in all models the effect estimate is positive as expected (i.e. higher temperature in May leads to more defoliation later in summer), the size of the estimate more

than triples when a spatial effect is added to the null model and the estimate in the spatial+ model is more than double that in the spatial model. This shows that, if we were to use the spatial model for our inference, the effect of temperature on crown defoliation would likely be underestimated in both size and significance due to spatial confounding. Note that, as expected, the gSEM gives similar results to spatial+.

6. Non-Gaussian response data

In this section we generalize the models from Section 2 to response distributions from the exponential family, which includes, e.g. the Poisson, gamma and binomial distributions.

6.1 *Spatial model*

Suppose we have response data $\mathbf{y} = (y_1, \ldots, y_n)^T$ where each y_i is assumed to be a random variable whose distribution is from the exponential family with $\mathrm{E}(y_i) = \mu_i$, and suppose $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{t}_1, \ldots, \mathbf{t}_n$ are covariate observations and spatial locations as before. A generalized version of (1) can then be formulated as

$$g(\mu_i) = \beta x_i + f(\mathbf{t}_i) \tag{13}$$

where β is an unknown parameter, f a thin plate spline and $g : \mathbb{R} \to \mathbb{R}$ a link function (i.e. a monotonic smooth function which ensures $g(\mu_i)$ is in the domain of the response variable). The partial thin plate spline estimates $\widehat{\beta}$ and $\widehat{\mathbf{f}} = (\widehat{f}(\mathbf{t}_1), \ldots, \widehat{f}(\mathbf{t}_n))^T$ are found using the penalized iterative re-weighted least squares (PIRLS) algorithm (for details, see Web Appendix G). If no smoothing is applied, these estimates are the maximum likelihood estimates in a generalized linear model (GLM), which are asymptotically unbiased.

6.2 *Spatial+ model*

Starting with the model (13), let $\mathbf{W}$ and $\mathbf{z}$ denote the weights matrix and pseudodata at convergence of the PIRLS algorithm. We then define the corresponding spatial+ model as follows. Let $\widehat{\mathbf{f}}^x$ and $\mathbf{r}^x = \mathbf{x} - \widehat{\mathbf{f}}^x = (r_1^x, \ldots, r_n^x)^T$ denote the fitted values and residuals in the

weighted thin plate regression (4) with weights $\mathbf{W}$, i.e. $\widehat{\mathbf{f}}^x$ is the minimizer of $\|\sqrt{\mathbf{W}}(\mathbf{x} - \mathbf{f}^x)\|^2 + n\lambda_x \mathbf{f}^{xT}\mathbf{\Gamma}\mathbf{f}^x$ with smoothing parameter $\lambda_x > 0$ and $\mathbf{\Gamma}$ defined as in (2). The spatial+ model is then the partial thin plate spline model defined by

$$g(\mu_i) = \beta r_i^x + f^+(\mathbf{t}_i) \tag{14}$$

where β and f^+ are estimated as in Section 6.1. For further details, see Web Appendix G.

6.3 *Simulations*

The models (13) and (14) can once again be implemented using thin plate regression splines in `mgcv`. To test the performance of the spatial+ model (14), we repeat the simulations from Section 4 for three different response distributions, namely, the Poisson distribution with canonical link function $g(\mu) = \log(\mu)$, the exponential distribution with (non-canonical) link function $g(\mu) = \log(\mu)$ and the binomial distribution with size parameter $n_{\text{bin}} = 10$ and canonical link function $g(\mu) = \log(\mu/(n_{\text{bin}} - \mu))$.

For each response distribution, we simulate 100 replicates of the response data $\mathbf{y} = (y_1, \ldots, y_n)^T$ by independently sampling each y_i from the given distribution with mean $\mu_i = g^{-1}(\eta_i)$ where $\eta_i = \beta x_i + f_i$ with $\mathbf{x} = (x_1, \ldots, x_n)^T$ simulated as in Section 4.1, $\sigma_x = 0.1$, and true effects $\beta = 3$, $\mathbf{f} = (f_1, \ldots, f_n)^T = -\mathbf{z} - \mathbf{z}'$ as before. The results of fitting the models (13) and (14) are summarized in Figure 3. For comparison we have also included the results of fitting the corresponding null model (i.e. the GLM defined by $g(\mu_i) = \beta x_i$) and the models (13) and (14) with no smoothing penalty applied. Finally, we have fitted a generalized version of the RSR model (for details see Web Appendix G). Note that we have not included the gSEM here as it is not immediately clear how to generalize this model to non-Gaussian response distributions.

We see that for all three response distributions, the overall behaviour of the models is similar to what we saw in the Gaussian case. As before, the null model and RSR model both have highly biased covariate effect estimates, however, note that unlike the Gaussian

case, the estimate is not the same in the two models. This is because, while in both models the estimate is given by $\widehat{\beta} = (\mathbf{x}^T\mathbf{W}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W}\mathbf{z}$, the fitted values, and hence the weights and pseudodata at convergence, differ. Without smoothing, as expected, the spatial and spatial+ models give the same results, however, while the covariate effect estimate looks unbiased for the Poisson and exponential response distributions, it looks slightly biased for the binomial distribution, though not materially. This is not surprising as GLMs are only asymptotically unbiased and may have some bias in practice, particularly, when the number of estimated parameters is relatively large as it is in this case (Cox and Snell, 1968). When smoothing is applied, MSE reduces as intended, but the covariate effect estimate in the spatial model becomes significantly biased while it remains broadly unbiased in the spatial+ model.

[Figure 3 about here.]

7. Discussion

We have shown that for covariates that are spatially dependent but not fully spatial, the proposed spatial+ model can be used to avoid unreliable effect estimates in spatial regression with clear advantages over existing methods. Our analysis also gives a clearer understanding of spatial confounding in this context. Spatial models, whether formulated in terms of spatially induced prior distributions or smoothing penalties, usually apply some form of spatial smoothing to reflect spatial correlation in the data and avoid overfitting. However, from the model formulation (1), we see that it is exactly this smoothing that causes spatial confounding bias as, without smoothing, the spatial model has unbiased estimates. The non-spatial information in the covariate means that the model can distinguish it from an unmeasured spatial confounder. However, if the correlation between the covariate and the spatial confounder is high, the smoothing applied to the spatial term in the model can disproportionately affect the estimate of the covariate effect.

The excessive smoothing-induced bias is avoided in both spatial+ and the gSEM. If no smoothing penalty is applied, both models give the same unbiased covariate effect estimates as the unsmoothed spatial model. Spatial+ reparametrizes the spatial model so that spatial dependence is removed from the covariates and instead fully contained in the spatial term f^+. This makes fixed effect estimates broadly independent of the spatial effects, in particular, they remain largely unbiased under spatial smoothing. The idea of decorrelating covariate and spatial terms is also used in RSR, however, restricting the spatial effects leads to bias by construction. In the gSEM, the elimination of all spatial information means that fixed effect estimates are once again decorrelated from the spatial effects and thereby protected from spatial smoothing. The resulting model of residuals only, however, seems less intuitive than spatial+ and, the change in response variable means that standard model selection criteria cannot be used for comparisons with the other models. A major advantage of spatial+ is also that the method generalizes easily to models with non-Gaussian response distributions and our simulations illustrate that the method still works well here.

Our above discussion shows that the decorrelation of effect estimates is the underlying reason why the spatial+ approach works. As mentioned in Section 1, the modification of the model matrix that achieves this is easily transferable to other spatial model formulations, and we would therefore expect the method to work well in general. However, as our theoretical derivations are specific to thin plate spline estimates, similar derivations or simulations could be done to confirm our results in other settings. One limitation to the spatial+ approach is that the covariate effects in the model must be linear. This assumption is needed for the spatial residuals to capture the true covariate effects. The spatial model (1) is easily extended, using the generalized additive model (GAM) framework, to include non-linear covariate terms in the form of smooths (i.e. unknown functions of the covariates estimated from the data). It would be interesting to see if any of the ideas of spatial+, as well as our

increased understanding of spatial confounding, can be used to develop methods for avoiding spatial confounding in this context.

Finally, applying spatial+ to the forestry example, we see that the effect of temperature on crown defoliation appears to be positive and significant as expected, and that this effect would likely be underestimated in both size and significance in the spatial model (and even more so in the null model). The other covariate, age of trees, in this example also illustrates that, if a covariate is not spatially confounded, this can be confirmed by showing that its effect estimate in the spatial and spatial+ models agree. It is possible that this idea could be used to develop a diagnostic or test that practitioners could use to identify spatial confounding in applications.

Acknowledgements

Emiko Dupont is supported by a scholarship from the EPSRC Centre for Doctoral Training in Statistical Applied Mathematics at Bath (SAMBa), under the project EP/L015684/1. We thank The Forest Institute Baden-Württemberg (Germany) for making the 2013 Terrestrial Crown Condition Inventory (TCCI) forest health monitoring survey data available.

References

Augustin, N., Musio, M., von Wilpert, K., Kublin, E., Wood, S., and Schumacher, M. (2009). Modeling spatiotemporal forest health monitoring data. *Journal of the American Statistical Association* **104,** 899–911.

Chen, H. and Shiau, J.-J. H. (1991). A two-stage spline smoothing method for partially linear models. *Journal of Statistical Planning and Inference* **27,** 187–201.

Chen, H. and Shiau, J.-J. H. (1994). Data-driven efficient estimators for a partially linear model. *The Annals of Statistics* pages 211–237.

Clayton, D. G., Bernardinelli, L., and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology* **22,** 1193–1202.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)* **30,** 248–265.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer.

Eichhorn, J., Roskams, P., Potočič, N., Timmermann, V., Ferretti, Mues, V., Szepesi, A., Durrant, D., Seletković, I., H-W.Schröck, Nevalainen, S., Bussotti, F., Garcia, P., and Wulff, S., editors (2017). *ICP Forests manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis of the effects of air pollution on forests.* Thünen Institute of Forest Ecosystems, Eberswalde,Germany.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a bayesian perspective. *Statistica Sinica* pages 731–761.

Hanks, E. M., Schliep, E. M., Hooten, M. B., and Hoeting, J. A. (2015). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics* **26,** 243–254.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician* **64,** 325–334.

Khan, K. and Calder, C. A. (2020). Restricted spatial regression methods: Implications for inference. *Journal of the American Statistical Association* pages 1–13.

Kimeldorf, G. S. and Wahba, G. (1970). Spline functions and stochastic processes. *Sankhyā: The Indian Journal of Statistics, Series A* pages 173–180.

Kirkwood, B. R. and Sterne, J. AC. (2010). Essential medical statistics. *John Wiley & Sons*

McNamee, R. (2003). Confounding and confounders. *Occupational and environmental medicine* **60 (3),** 227–234.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25,** 107.

Page, G. L., Liu, Y., He, Z., and Sun, D. (2017). Estimation and prediction in the presence of spatial confounding for spatial linear models. *Scandinavian Journal of Statistics* .

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* **62,** 1197–1206.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics & probability letters* **4,** 203–208.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)* **47,** 1–21.

Thaden, H. and Kneib, T. (2018). Structural equation models for dealing with spatial confounding. *The American Statistician* **72,** 239–252.

Utreras, F. I. (1988). Convergence rates for multivariate smoothing spline functions. *Journal of approximation theory* **52,** 1–27.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Supporting Information

Web Appendices referenced in Sections 1, 3, 4 and 6, along with the `R` code for Sections 4, 5 and 6 are available with this paper at the Biometrics website on Wiley Online Library.

Received July 2020. *Revised July* 2020. *Accepted July* 2020.

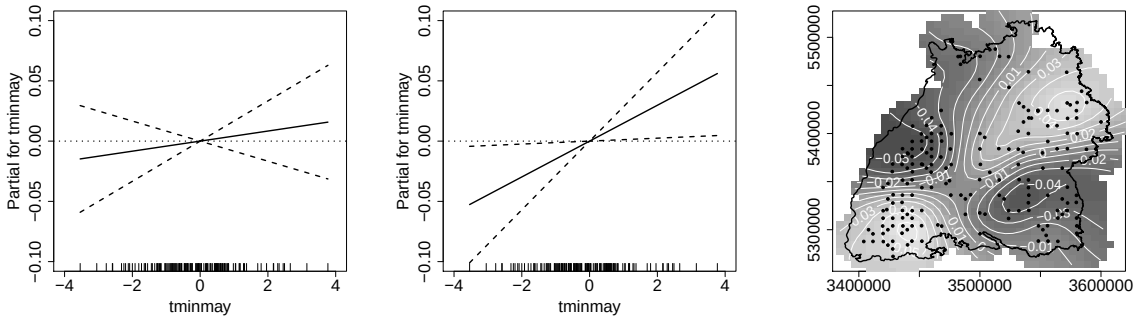


Figure 1. Forestry example. Estimated effect of minimum temperature in May on crown defoliation in the null model (left) and the spatial model (middle), where for each model the plot shows the contribution of the centered covariate to the predicted response (with two times standard error bands). Estimated spatial effect in the spatial model (right) with the border of Baden-Württemberg outlined and dots showing the data locations.

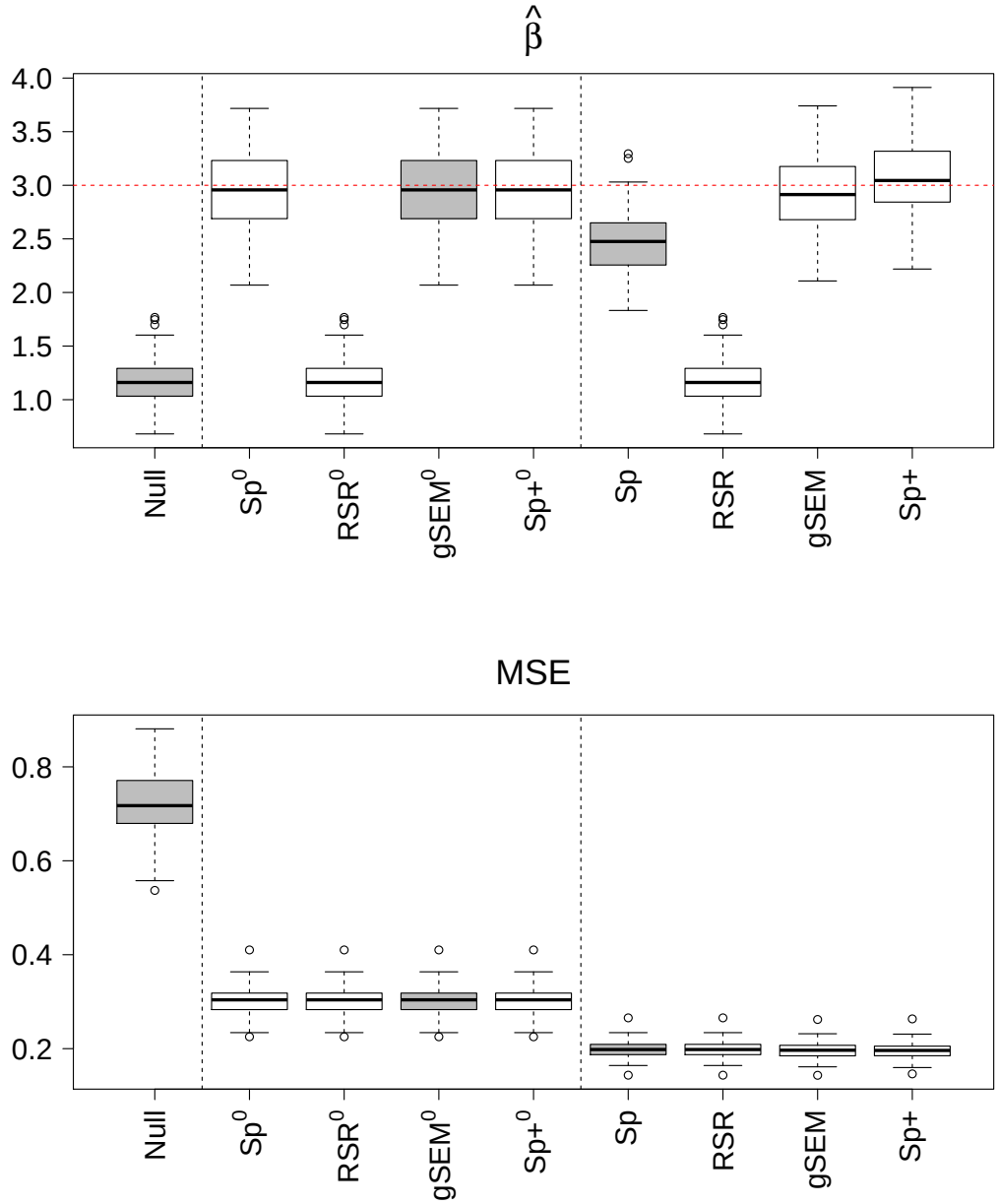


Figure 2. Estimated covariate effect $\widehat{\beta}$ (top) and MSE of fitted values (bottom) for each model fitted to 100 data replicates, where the true covariate effect is $\beta = 3$. Sp and Sp+ denote the spatial and spatial+ models, respectively, and superscript 0 refers to an unsmoothed model (i.e. $\lambda = \lambda_x = 0$). Results in grey are the three models that correspond to those used in Thaden and Kneib's simulation study.

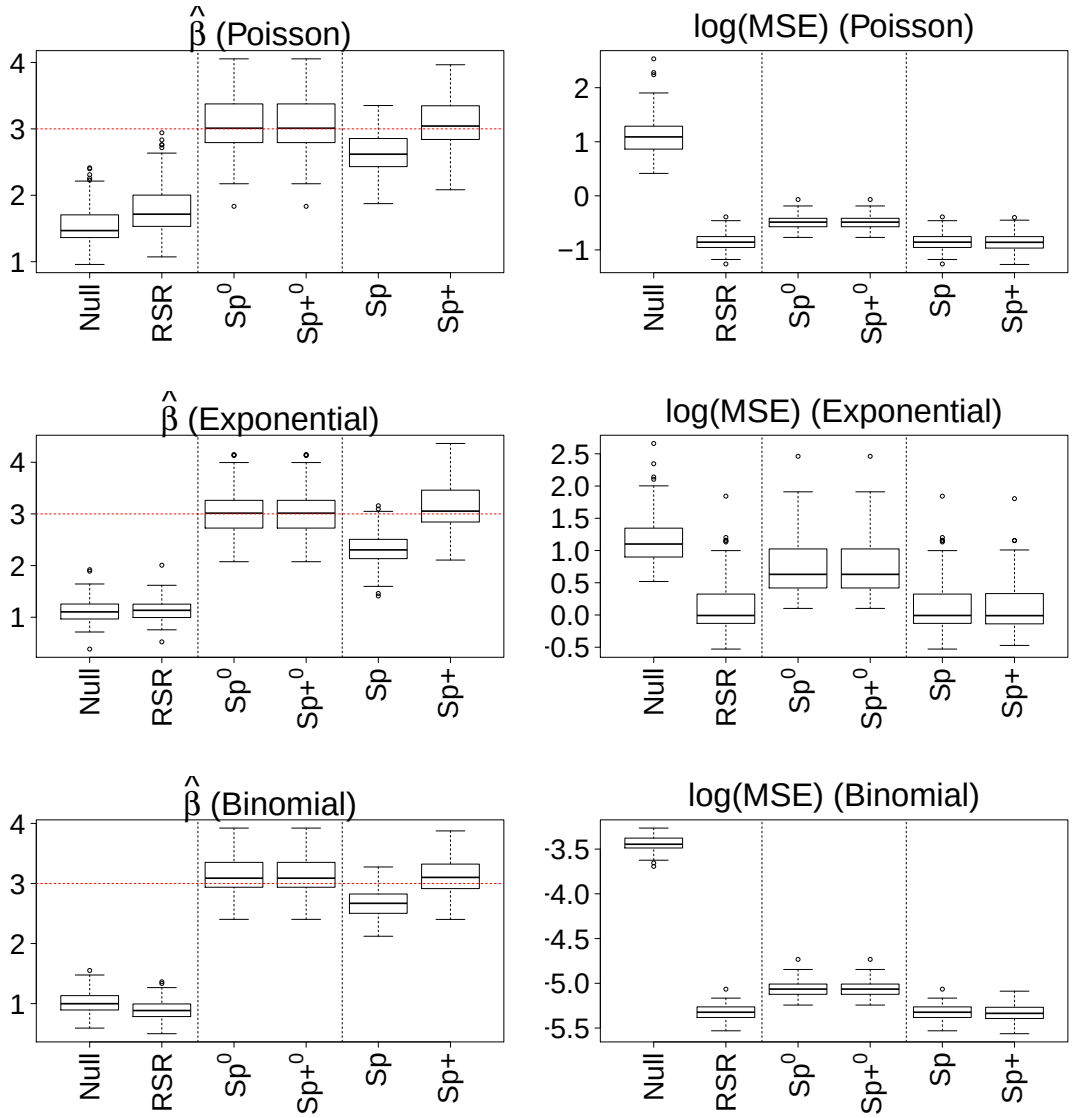


Figure 3. For each of the distributions Poisson (top), exponential (middle), binomial (bottom): the estimated covariate effect $\widehat{\beta}$ (left) and log(MSE) of fitted values (right) for each model fitted to 100 data replicates, where the true covariate effect is $\beta = 3$. Sp and Sp+ denote the spatial and spatial+ models, respectively, and superscript 0 refers to an unsmoothed model (i.e. $\lambda = \lambda_x = 0$).

Table 1

Forestry example: results of fitting models to the data. For each covariate: the estimate of the covariate effect β and its p-value. `s(x,y)` *refers to the thin plate regression splines fitted to f in the spatial model and f^+ in the spatial+ model. For each of these: the effective degrees of freedom (edf) and the p-value. For each significant p-value we write *** if it is < 0.001, ** if < 0.01 and * if < 0.05. Note that in the gSEM, deviance explained, estimated standard deviation and AIC do not compare directly with the other models as the response variable is different.*

	`age`			`tminmay`			`s(x,y)`					
	$\widehat{\beta}$	p-value		$\widehat{\beta}$	p-value		edf	p-value		Dev expl	$\widehat{\sigma}$	AIC
Null	0.00247	$< 10^{-16}$	***	0.0042	0.5049					0.490	0.00940	-335
Spatial	0.00237	$< 10^{-16}$	***	0.0149	0.0307	*	14.2	0.0243	*	0.605	0.00789	-355
Spatial+	0.00237	$< 10^{-16}$	***	0.0316	0.0073	**	12.0	3.32e-05	***	0.598	0.00793	-356
gSEM	0.00232	$< 10^{-16}$	***	0.0317	0.0058	**						

Supporting information for "Spatial+: a novel approach to spatial confounding" by Emiko Dupont, Simon N. Wood and Nicole H. Augustin

1 Web Appendix A: Technical lemmas

In this appendix we set out the technical lemmas that we use for the derivations of the main results of our asymptotic analysis (detailed in Web Appendix C below), which generalize the results of Rice (1986); Chen and Shiau (1991) from $d = 1$ to dimensions $d \geq 1$. Key to this generalization is the following result by Utreras (1988) on the asymptotics of thin plate splines.

Lemma 1.1. *Suppose Ω has Lipschitz boundary and satisfies a uniform cone condition (as defined in Utreras (1988)). Assume that the points $\{\mathbf{t}_1, \ldots, \mathbf{t}_n\} \subset \Omega$ are regularly distributed in the sense that there exists a constant $B > 0$ such that*

$$\frac{h_{\min}}{h_{\max}} \leq B$$

where $h_{\max} = \sup_{\mathbf{t}\in\Omega} \inf_i |\mathbf{t} - \mathbf{t}_i|$ and $h_{\min} = \min_{i\neq j} |\mathbf{t}_i - \mathbf{t}_j|$. Let $\mu_1 \leq \cdots \leq \mu_n$ denote the eigenvalues of the matrix $n\mathbf{\Gamma}$ and assume $m > d/2$. Then

$$\mu_1 = \cdots = \mu_M = 0$$

and there exist constants $C_1, C_2 > 0$ such that

$$C_1 k^{2m/d} \leq \mu_k \leq C_2 k^{2m/d} \quad \textit{for } M + 1 \leq k \leq n.$$

Proof. See the proof of Theorem 5.1 (a) and Theorem 5.3 of Utreras (1988). □

Lemma 1.1 provides us with a convenient basis in which the smoother matrix $\mathbf{S}_\lambda = (\mathbf{I} + n\lambda\mathbf{\Gamma})^{-1}$ is diagonalized and, moreover, describes the asymptotic behaviour of its eigenvalues as the number of data points $n \to \infty$. More specifically, if $\mathbf{\Phi}$ is the matrix whose columns are $\frac{1}{\sqrt{n}}\boldsymbol{\phi}_1, \ldots, \frac{1}{\sqrt{n}}\boldsymbol{\phi}_n$ where $\boldsymbol{\phi}_k$ is an eigenvector of $n\Gamma$ corresponding to the eigenvalue μ_k, then (with appropriate scaling of the eigenvectors) $\mathbf{\Phi}$ has orthonormal columns and

$$\begin{aligned} \mathbf{\Phi}^T\mathbf{S}_\lambda\mathbf{\Phi} &= \operatorname{diag}\big(1/(1+\lambda\mu_1), \ldots, 1/(1+\lambda\mu_n)\big), \\ \mathbf{\Phi}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{\Phi} &= \operatorname{diag}\big((\lambda\mu_1)/(1+\lambda\mu_1), \ldots, (\lambda\mu_n)/(1+\lambda\mu_n)\big). \end{aligned}$$

This representation allows us to explicitly evaluate the estimates in the models of dimension $d \geq 1$ which, in turn, enables us to obtain asymptotic results in a similar way to Rice (1986); Chen and Shiau (1991).

For the rest of the supplementary web material, we assume that $m > d/2$ and that the domain Ω and the data points $\mathbf{t}_1, \ldots, \mathbf{t}_n$ satisfy the conditions of Lemma 1.1. We will also use the notation $a(n) \approx b(n)$ to mean that $a(n)/b(n)$ is bounded away from zero and infinity as $n \to \infty$.

Lemmas 1.2 and 1.3 link the asymptotic behaviour of the smoother matrix $\mathbf{S}_\lambda$ to the convergence rate of the smoothing parameter λ. Lemma 1.2 generalizes Lemma 2 of Chen and Shiau (1991) to dimensions $d \geq 1$, and is proved using the asymptotic properties of the eigenvalues given in Lemma 1.1. The result in

Lemma 1.3 is proved by Utreras (1988). Lemmas 1.4 and 1.5 prove a number of asymptotic results that are convenient for later proofs. Lemma 1.4 shows how the results used by Rice (1986) for the analysis in dimension $d = 1$ generalize to dimensions $d \geq 1$, while Lemma 1.5 generalizes Lemma 3 of Chen and Shiau (1991) to dimensions $d \geq 1$. Proofs of Lemmas 1.2, 1.4 and 1.5 are given in Web Appendix B.

Lemma 1.2. *Suppose $\lambda \approx n^{-\delta}$ for some $0 < \delta < 1$. Then*

(a) $Tr(\mathbf{S}_\lambda) = \sum_{k=1}^{n}(1+\lambda\mu_k)^{-1} = M + \mathcal{O}(\lambda^{-d/2m})$,

(b) $Tr(\mathbf{S}_\lambda^2) = \sum_{k=1}^{n}(1+\lambda\mu_k)^{-2} = M + \mathcal{O}(\lambda^{-d/2m})$.

In particular, if $m \geq d$, then both of these sums are of the form $\mathcal{O}(n^{1/2-\tau})$ where $0 < \tau < 1/2$ depends only on δ.

Proof. See Web Appendix B. □

Lemma 1.3. *For any $g \in H^m(\Omega)$, let $\mathbf{g} = (g(\mathbf{t}_1), \ldots, g(\mathbf{t}_n))^T$. The averaged squared bias $B^2_{\text{tp}}(g,\lambda)$ of the thin plate spline $\mathbf{S}_\lambda\mathbf{g}$ (i.e. the fitted values in a model of the form (1) in our paper with $\beta = 0$) is given by*

$$B^2_{\text{tp}}(g,\lambda) = \frac{1}{n}\mathbf{g}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{g} = \mathcal{O}(\lambda).$$

Proof. See Utreras (1988) Lemma 2.2. □

Lemma 1.4. *Suppose $\lambda \approx n^{-\delta}$ for some $0 < \delta < 1$, $f, f^x \in H^m(\Omega)$ are bounded and $m \geq d$. Let $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$. Then*

(a) $n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{x} = \sigma_x^2 + o(1)$,

(b) $n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{x} = \sigma_x^2 + o(1)$,

(c) $n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{f} = o(n^{-1/2}) + \mathcal{O}(\lambda^{-1/2})$,

(d) $n^{-1}\mathbf{x}^T\mathbf{S}_\lambda^2\mathbf{x} = \mathcal{O}(1)$

Proof. See Web Appendix B. □

Lemma 1.5. *Suppose $\lambda \approx n^{-\delta}, \lambda_x \approx n^{-\delta_x}$ for some $0 < \delta, \delta_x < 1$, $f, f^x \in H^m(\Omega)$ and $m \geq d$. Let $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$. Then*

(a) $n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_{\lambda_x})(\mathbf{I}-\mathbf{S}_\lambda)(\mathbf{I}-\mathbf{S}_{\lambda_x})\mathbf{x} = \sigma_x^2 + o(1)$,

(b) $n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_{\lambda_x})(\mathbf{I}-\mathbf{S}_\lambda)^2(\mathbf{I}-\mathbf{S}_{\lambda_x})\mathbf{x} = \sigma_x^2 + o(1)$,

(c) $n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_{\lambda_x})(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{f} = o(n^{-1/2}) + \mathcal{O}((\lambda_x\lambda)^{1/2})$,

(d) $n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_{\lambda_x})(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\mathbf{x} = o(n^{-1/2}) + \mathcal{O}((\lambda_x\lambda)^{1/2})$,

(e) $n^{-1}\mathbf{x}^T\mathbf{S}_{\lambda_x}(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{S}_{\lambda_x}\mathbf{x} = \mathcal{O}(\lambda) + \mathcal{O}(n^{-1}\lambda_x^{-d/2m}\log^2 n)$

(f) $n^{-1}\mathbf{x}^T[\mathbf{S}_\lambda + (\mathbf{I}-\mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}]^T[\mathbf{S}_\lambda + (\mathbf{I}-\mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}]\mathbf{x} = \mathcal{O}(1)$,

Proof. See Web Appendix B. □

2 Web Appendix B: Proofs of technical lemmas

In this appendix we prove the lemmas set out in Web Appendix A. We start by introducing some notation. Recall the assumption from our paper that

$$x_i = f^x(\mathbf{t}_i) + \epsilon_i^x, \quad \epsilon_i^x \sim_{\text{iid}} N(0, \sigma_x^2)$$

which means that the covariate $\mathbf{x}$ is correlated with the smooth f in the spatial model. Therefore, $\mathbf{x}$ decomposes as

$$\mathbf{x} = \mathbf{f}^x + \boldsymbol{\epsilon}^x \tag{1}$$

with $\mathbf{f}^x = (f^x(\mathbf{t}_1), \ldots, f^x(\mathbf{t}_n))^T$ and $\boldsymbol{\epsilon}^x = (\epsilon_1^x, \ldots \epsilon_n^x)^T$. For the asymptotic analysis, it is often convenient to consider the behaviour of the components in this decomposition separately. Let $\mathbf{c}^x = (c_1^x, \ldots, c_n^x)^T$ and $\boldsymbol{\xi}^x = (\xi_1^x, \ldots, \xi_n^x)^T$ denote the coefficients of $\mathbf{f}^x$ and $\boldsymbol{\epsilon}^x$, respectively, in the basis $\boldsymbol{\Phi}$ introduced in Web Appendix A, i.e.

$$\begin{aligned} \mathbf{f}^x &= \boldsymbol{\Phi}\mathbf{c}^x \quad \text{where } \mathbf{c}^x = \boldsymbol{\Phi}^T\mathbf{f}^x, \\ \boldsymbol{\epsilon}^x &= \boldsymbol{\Phi}\boldsymbol{\xi}^x \quad \text{where } \boldsymbol{\xi}^x = \boldsymbol{\Phi}^T\boldsymbol{\epsilon}^x. \end{aligned}$$

Note that since $f^x \in H^m(\Omega)$ is bounded, we have that

$$n^{-1}\sum_{k=1}^{n}(c_k^x)^2 = n^{-1}(\mathbf{f}^x)^T(\mathbf{f}^x) \to 0 \text{ as } n \to \infty. \tag{2}$$

As in Rice (1986) and Chen and Shiau (1991), we also note that the following assumptions hold for the coefficients $\boldsymbol{\xi}^x$ of the iid noise $\boldsymbol{\epsilon}^x$.

(A1) $n^{-1}\sum_{k=1}^{n}\xi_k^x \to 0$ as $n \to \infty$,

(A2) $n^{-1}\sum_{k=1}^{n}(\xi_k^x)^2 = n^{-1}(\boldsymbol{\epsilon}^x)^T\boldsymbol{\epsilon}^x \to \sigma_x^2 > 0$ as $n \to \infty$,

(A3) $\sup_{1\le k\le n}|\xi_k^x| = \mathcal{O}(\log n)$.

Proof of Lemma 1.2

From Lemma 1.1, $\mu_k = 0$ for $k = 1, \ldots, M$, so $\sum_{k=1}^{M}(1+\lambda\mu_k)^{-1} = M$. Split the remaining range of the summation into $I_1 = [M+1, \lambda^{-d/2m}]$, $I_2 = [\lambda^{-d/2m}, n]$.

I_1: Since $(1+\lambda\mu_k)^{-1} \le 1$ for all k

$$\sum_{I_1}(1+\lambda\mu_k)^{-1} \le \sum_{I_1} 1 \le \lambda^{-d/2m}.$$

I_2: By Lemma 1.1, $(1+\lambda\mu_k)^{-1} \le (C_1\lambda k^{2m/d})^{-1}$ for all k in I_2. Since $\{\mu_k\}_k$ is an increasing sequence, we have that

$$\begin{aligned} \sum_{I_2}(1+\lambda\mu_k)^{-1} &\le \int_{\lambda^{-d/2m}}^{\infty}(C_1\lambda x^{2m/d})^{-1}dx \\ &= C\lambda^{-d/2m} \end{aligned}$$

where $C = (C_1(2m/d-1))^{-1}$. This proves part (a).

For part (b) we note that $\sum_{k=1}^{M}(1+\lambda\mu_k)^{-2} = M$ as before and that $(1+\lambda\mu_k)^{-2} < (1+\lambda\mu_k)^{-1}$ for all the remaining k. Therefore (b) follows from (a).

Proof of Lemma 1.4

To prove (a), we use the decomposition $\mathbf{x} = \mathbf{f}^x + \boldsymbol{\epsilon}^x$ from (1) and the corresponding basis expansions in the basis $\boldsymbol{\Phi}$ to get

$$n^{-1}\mathbf{x}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{x} = n^{-1}\sum_k (c_k^x + \xi_k^x)^2 \frac{\lambda\mu_k}{1+\lambda\mu_k}.$$

We note that while

$$(c_k^x + \xi_k^x)^2 = (c_k^x)^2 + (\xi_k^x)^2 + 2c_k^x\xi_k^x,$$

due to the Cauchy-Schwarz inequality, the term $2c_k^x\xi_k^x$ will never dominate the rate of convergence. Therefore, we only need to consider the parts of the sum relating to the other two terms. Using Cauchy-Schwarz again we see that

$$\begin{aligned}
\sum_k (c_k^x)^2 \frac{\lambda\mu_k}{1+\lambda\mu_k} &\leq \Big(\sum_k (c_k^x)^2 \big(\frac{\lambda\mu_k}{1+\lambda\mu_k}\big)^2\Big)^{1/2} \Big(\sum_k (c_k^x)^2\Big)^{1/2} \\
&= \big(nB_{\text{tp}}^2(f^x,\lambda)\big)^{1/2}\big(\sum_k (c_k^x)^2\big)^{1/2} \\
&= \mathcal{O}(n\lambda^{1/2}) = \mathcal{O}(n^{1-\delta/2}) = o(n).
\end{aligned}$$

Here we have used Lemma 1.3 and (2).

For the term involving $(\xi_k^x)^2$ we have that

$$\begin{aligned}
\sum_k (\xi_k^x)^2 - \sum_k (\xi_k^x)^2 \frac{\lambda\mu_k}{1+\lambda\mu_k} &= \sum_k (\xi_k^x)^2 \frac{1}{1+\lambda\mu_k} \\
&\leq \sup_k (\xi_k^x)^2 \sum_k \frac{1}{1+\lambda\mu_k} \\
&= \mathcal{O}(\log^2 n)\mathcal{O}(n^{1/2-\tau}) = o(n)
\end{aligned}$$

by assumption (A3) and Lemma 1.2. Hence, by assumption (A2),

$$n^{-1}\sum_k (\xi_k^x)^2 \frac{\lambda\mu_k}{1+\lambda\mu_k} \to \sigma_x^2 \quad \text{as } n \to \infty,$$

and therefore (a) is proved.

For (b) we write

$$n^{-1}\mathbf{x}^T(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{x} = n^{-1}\sum_k (c_k^x + \xi_k^x)^2 \Big(\frac{\lambda\mu_k}{1+\lambda\mu_k}\Big)^2.$$

By Lemma 1.3 we have that

$$n^{-1}\sum_k (c_k^x)^2 \Big(\frac{\lambda\mu_k}{1+\lambda\mu_k}\Big)^2 = B_{\text{tp}}^2(f^x,\lambda) = \mathcal{O}(\lambda) = o(1).$$

For $a > 0$ we have $\frac{1}{1+a} \leq 1$ and $\frac{a}{1+a} \leq 1$ and therefore

$$1 - \Big(\frac{a}{1+a}\Big)^2 = \frac{(1+a)^2 - a^2}{(1+a)^2} = \frac{(1+a)+a}{(1+a)^2} \leq \frac{2}{1+a}.$$

Using this with $a = \lambda\mu_k$ we see from assumption (A3) and Lemma 1.2 that

$$\begin{aligned}
\sum_k (\xi_k^x)^2 - \sum_k (\xi_k^x)^2 \Big(\frac{\lambda\mu_k}{1+\lambda\mu_k}\Big)^2 &\leq \sup_k (\xi_k^x)^2 \sum_k \frac{2}{1+\lambda\mu_k} \\
&= \mathcal{O}\big((\log^2 n)n^{1/2-\tau}\big) = o(n).
\end{aligned}$$

So by assumption (A2), (b) is proved.

For (c) let $\mathbf{c} = \mathbf{\Phi}^T\mathbf{f}$ be the coefficients of $\mathbf{f}$ in the basis $\mathbf{\Phi}$. Then

$$n^{-1}\mathbf{x}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{f} = n^{-1}\sum_k (c_k^x c_k + \xi_k^x c_k)\frac{\lambda\mu_k}{1+\lambda\mu_k}.$$

For the term involving c_k^x, we use Cauchy-Schwarz and (2) to see that

$$\begin{aligned}\left|n^{-1}\sum_k c_k^x c_k \frac{\lambda\mu_k}{1+\lambda\mu_k}\right| &\leq \left(n^{-1}\sum_k (c_k^x)^2\right)^{1/2}\left(n^{-1}\sum_k\left(\frac{c_k\lambda\mu_k}{1+\lambda\mu_k}\right)^2\right)^{1/2} \\ &= \mathcal{O}\big((B_{\text{tp}}^2(f,\lambda)\big)^{1/2}) = \mathcal{O}(\lambda^{1/2})\end{aligned}$$

by Lemma 1.3. For the term involving ξ_k^x, we use Cauchy-Schwarz again to obtain

$$\begin{aligned}\left|n^{-1}\sum_k \xi_k^x c_k \frac{\lambda\mu_k}{1+\lambda\mu_k}\right| &\leq \lambda^{1/2}\sup_k|\xi_k^x|\left|n^{-1}\sum_k c_k\mu_k^{1/2}\frac{(\lambda\mu_k)^{1/2}}{1+\lambda\mu_k}\right| \\ &\leq \lambda^{1/2}\sup_k|\xi_k^x|\left(n^{-1}\sum_k c_k^2\mu_k\right)^{1/2}\left(n^{-1}\sum_k\frac{\lambda\mu_k}{(1+\lambda\mu_k)^2}\right)^{1/2} \\ &\leq \mathcal{O}(\lambda^{1/2}\log n)\mathcal{O}(n^{-1/2}\lambda^{-d/4m}) = o(n^{-1/2})\end{aligned}$$

Here we have used assumption (A3), Lemma 1.2 (since $\frac{\lambda\mu_k}{(1+\lambda\mu_k)^2} \leq \frac{1}{1+\lambda\mu_k}$) and the fact that

$$n^{-1}\sum_k c_k^2\mu_k = \mathbf{f}^T\Gamma\mathbf{f} \leq |f|_m^2 < \infty$$

since $f \in H^m(\Omega)$. The rate of convergence of $o(n^{-1/2})$ follows from the fact that

$$n^{-1/2}(\log n)\lambda^{-d/4m+1/2} \approx n^{-1/2}(\log n)n^{-\delta(1-d/2m)/2} = o(n^{-1/2})$$

since $1 - d/2m > 0$. This proves (c).

For (d) we have that

$$n^{-1}\mathbf{x}^T\mathbf{S}_\lambda^2\mathbf{x} = n^1\sum_k (c_k^x + \xi_k^x)^2\frac{1}{(1+\lambda_x\mu_k)^2}.$$

For the term involving $(c_k^x)^2$ we see that

$$n^{-1}\sum_k (c_k^x)^2\frac{1}{(1+\lambda_x\mu_k)^2} \leq n^{-1}\sum_k (c_k^x)^2 = \mathcal{O}(1)$$

by (2). For the term involving $(\xi_k^x)^2$ we see from assumption (A3) and Lemma 1.2 that

$$\begin{aligned}n^{-1}\sum_k (\xi_k^x)^2\frac{1}{(1+\lambda_x\mu_k)^2} &\leq n^{-1}\sup_k(\xi_k^x)^2\sum_k\frac{1}{(1+\lambda_x\mu_k)^2} \\ &= \mathcal{O}\big((\log^2 n)n^{-1/2-\tau}\big) = \mathcal{O}(1).\end{aligned}$$

Hence $n^{-1}\mathbf{x}^T\mathbf{S}_\lambda^2\mathbf{x} = \mathcal{O}(1)$.

Proof of Lemma 1.5

As in the proof of Lemma 1.4 we write

$$n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_{\lambda_x})(\mathbf{I}-\mathbf{S}_{\lambda})(\mathbf{I}-\mathbf{S}_{\lambda_x})\mathbf{x} = n^{-1}\sum_k (c_k^x+\xi_k^x)^2\left(\frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\right)^2\frac{\lambda\mu_k}{1+\lambda\mu_k}$$

and once again, by Cauchy-Schwarz, we only need to consider the terms involving $(c_k^x)^2$ and $(\xi_k^x)^2$. Since $\frac{\lambda\mu_k}{1+\lambda\mu_k}\leq 1$, Lemma 1.3 shows that

$$n^{-1}\sum_k (c_k^x)^2\left(\frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\right)^2\frac{\lambda\mu_k}{1+\lambda\mu_k} \leq n^{-1}(\mathbf{f}^x)^T(\mathbf{I}-\mathbf{S}_{\lambda_x})^2\mathbf{f}^x = B_{\mathrm{tp}}^2(f^x,\lambda_x) = \mathcal{O}(\lambda_x) = o(1).$$

For the term involving $(\xi_k^x)^2$, firstly note that if $a_1,a_2,a_3>0$, then

$$\begin{aligned}
1-\frac{a_1a_2a_3}{(1+a_1)(1+a_2)(1+a_3)} &= \frac{(1+a_1)(1+a_2)(1+a_3)-a_1a_2a_3}{(1+a_1)(1+a_2)(1+a_3)}\\
&= \frac{1+a_1+a_2+a_3+a_1a_2+a_1a_3+a_2a_3}{(1+a_1)(1+a_2)(1+a_3)}\\
&\leq \frac{3}{1+a_1}+\frac{2}{1+a_2}+\frac{2}{1+a_3}
\end{aligned}$$

where in the last step we have used the fact that $\frac{1}{1+a_i}\leq 1$ and $\frac{a_i}{1+a_i}\leq 1$ for all i. Using this with $a_1=a_2=\lambda_x\mu_k$ and $a_3=\lambda\mu_k$ we see that

$$\begin{aligned}
\sum_k(\xi_k^x)^2-\sum_k(\xi_k^x)^2\left(\frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\right)^2\frac{\lambda\mu_k}{1+\lambda\mu_k} &\leq \sup_k(\xi_k^x)^2\left(\sum_k\frac{5}{1+\lambda_x\mu_k}+\sum_k\frac{2}{1+\lambda_x\mu_k}\right)\\
&= \mathcal{O}(\log^2 n)\mathcal{O}(n^{1/2-\tau}) = o(n)
\end{aligned}$$

by assumption (A3) and Lemma 1.2. Therefore,

$$n^{-1}\sum_k(\xi_k^x)^2\left(\frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\right)^2\frac{\lambda\mu_k}{1+\lambda\mu_k}\to\sigma_x^2$$

by assumption (A2). This shows (a).

For (b) we have that

$$n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_{\lambda_x})(\mathbf{I}-\mathbf{S}_{\lambda})^2(\mathbf{I}-\mathbf{S}_{\lambda_x})\mathbf{x} = n^{-1}\sum_k(c_k^x+\xi_k^x)^2\left(\frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\right)^2\left(\frac{\lambda\mu_k}{1+\lambda\mu_k}\right)^2.$$

For the term involving $(c_k^x)^2$, the same argument as in (a) shows that this is $o(1)$. For the $(\xi_k^x)^2$ term we note that

$$1-\frac{a_1a_2a_3a_4}{(1+a_1)(1+a_2)(1+a_3)(1+a_4)}\leq\frac{5}{1+a_1}+\frac{4}{1+a_2}+\frac{4}{1+a_3}+\frac{2}{1+a_4}$$

for $a_1,a_2,a_3,a_4>0$ and using this with $a_1=a_2=\lambda_x\mu_k$ and $a_3=a_4=\lambda\mu_k$ shows that

$$n^{-1}\sum_k(\xi_k^x)^2\left(\frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\right)^2\left(\frac{\lambda\mu_k}{1+\lambda\mu_k}\right)^2\to\sigma_x^2$$

as in (a). This proves (b).

For (c) let $\mathbf{c} = \mathbf{\Phi}^T\mathbf{f}$ be the coefficients of $\mathbf{f}$ in the basis $\mathbf{\Phi}$. Then

$$n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_{\lambda_x})(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{f} = n^{-1}\sum_k (c_k^x c_k + \xi_k^x c_k)\frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\frac{\lambda\mu_k}{1+\lambda\mu_k}.$$

For the term involving c_k^x, we use Cauchy-Schwarz to see that

$$\begin{aligned}\left|n^{-1}\sum_k c_k^x c_k \frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\frac{\lambda\mu_k}{1+\lambda\mu_k}\right| &\le \left(n^{-1}\sum_k\left(\frac{c_k^x\lambda_x\mu_k}{1+\lambda_x\mu_k}\right)^2\right)^{1/2}\left(n^{-1}\sum_k\left(\frac{c_k\lambda\mu_k}{1+\lambda\mu_k}\right)^2\right)^{1/2}\\ &= \left(B_{\mathrm{tp}}^2(f^x,\lambda_x)B_{\mathrm{tp}}^2(f,\lambda)\right)^{1/2} = \mathcal{O}\big((\lambda_x\lambda)^{1/2}\big)\end{aligned}$$

by Lemma 1.3. For the term involving ξ_k^x, since $\frac{\lambda_x\mu_k}{1+\lambda_x\mu_k} \le 1$,

$$\left|n^{-1}\sum_k \xi_k^x c_k \frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\frac{\lambda\mu_k}{1+\lambda\mu_k}\right| \le \left|n^{-1}\sum_k \xi_k^x c_k \frac{\lambda\mu_k}{1+\lambda\mu_k}\right| = o(n^{-1/2})$$

by the proof of Lemma 1.4 (c). This proves (c).

For (d) we have that

$$n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_{\lambda_x})(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\mathbf{x} = n^{-1}\sum_k (c_k^x+\xi_k^x)^2\frac{\lambda_x\mu_k}{(1+\lambda_x\mu_k)^2}\frac{\lambda\mu_k}{1+\lambda\mu_k}.$$

For the term involving $(c_k^x)^2$, Cauchy-Schwarz implies that

$$\begin{aligned}n^{-1}\sum_k (c_k^x)^2\frac{\lambda_x\mu_k}{(1+\lambda_x\mu_k)^2}\frac{\lambda\mu_k}{1+\lambda\mu_k} &\le n^{-1}\sum_k (c_k^x)^2\frac{\lambda_x\mu_k}{1+\lambda_x\mu_k}\frac{\lambda\mu_k}{1+\lambda\mu_k}\\ &\le \big(B_{\mathrm{tp}}^2(f^x,\lambda_x)B_{\mathrm{tp}}^2(f^x,\lambda)\big)^{1/2} = \mathcal{O}\big((\lambda\lambda_x)^{1/2}\big)\end{aligned}$$

by Lemma 1.3. For the term involving $(\xi_k^x)^2$ we use (A3) and Lemma 1.2 to see that

$$\begin{aligned}n^{-1}\sum_k (\xi_k^x)^2\frac{\lambda_x\mu_k}{(1+\lambda_x\mu_k)^2}\frac{\lambda\mu_k}{1+\lambda\mu_k} &\le \sup_k(\xi_k^x)^2 n^{-1}\sum_k\frac{1}{1+\lambda_x\mu_k}\\ &= \mathcal{O}\big((\log^2 n)n^{-1/2-\tau}\big) = o(n^{-1/2}).\end{aligned}$$

This proves (d)

For (e) we have that

$$n^{-1}\mathbf{x}^T\mathbf{S}_{\lambda_x}(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{S}_{\lambda_x}\mathbf{x} = n^{-1}\sum_k (c_k^x+\xi_k^x)^2\frac{1}{(1+\lambda_x\mu_k)^2}\left(\frac{\lambda\mu_k}{1+\lambda\mu_k}\right)^2.$$

For the term involving $(c_k^x)^2$ we see that

$$\begin{aligned}n^{-1}\sum_k (c_k^x)^2\frac{1}{(1+\lambda_x\mu_k)^2}\left(\frac{\lambda\mu_k}{1+\lambda\mu_k}\right)^2 &\le n^{-1}\sum_k (c_k^x)^2\left(\frac{\lambda\mu_k}{1+\lambda\mu_k}\right)^2\\ &= B_{\mathrm{tp}}^2(f^x,\lambda) = \mathcal{O}(\lambda).\end{aligned}$$

For the term involving $(\xi_k^x)^2$

$$\begin{aligned}n^{-1}\sum_k (\xi_k^x)^2\frac{1}{(1+\lambda_x\mu_k)^2}\left(\frac{\lambda\mu_k}{1+\lambda\mu_k}\right)^2 &\le n^{-1}\sup_k(\xi_k^x)^2\sum_k\frac{1}{(1+\lambda_x\mu_k)^2}\\ &= \mathcal{O}\big(n^{-1}(\log^2 n)\lambda_x^{-d/2m}\big)\end{aligned}$$

by assumption (A3) and Lemma 1.2. This proves (e).

For (f) we write

$$\begin{aligned} & n^{-1}\mathbf{x}^T\big[\mathbf{S}_\lambda + (\mathbf{I}-\mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}]^T[\mathbf{S}_\lambda + (\mathbf{I}-\mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\big]\mathbf{x} \\ & = n^{-1}\big(\mathbf{x}^T\mathbf{S}_\lambda^2\mathbf{x} + 2\mathbf{x}^T\mathbf{S}_\lambda(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\mathbf{x} + \mathbf{x}^T\mathbf{S}_{\lambda_x}(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{S}_{\lambda_x}\big). \end{aligned} \tag{3}$$

For the first term in (3), $n^{-1}\mathbf{x}^T\mathbf{S}_\lambda^2\mathbf{x} = \mathcal{O}(1)$ by Lemma 1.4 (d). For the second term in (3) we see that

$$\begin{aligned} n^{-1}\mathbf{x}^T\mathbf{S}_\lambda(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\mathbf{x} &= n^{-1}\sum_k (c_k^x+\xi_k^x)^2 \frac{1}{1+\lambda_x\mu_k}\frac{\lambda\mu_k}{(1+\lambda\mu_k)^2} \\ &\le n^{-1}\sum_k (c_k^x+\xi_k^x)^2 \frac{\lambda\mu_k}{1+\lambda\mu_k} = n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{x} = \mathcal{O}(1) \end{aligned}$$

by Lemma 1.4 (a). From (e), the third term in (3) is given by

$$\begin{aligned} n^{-1}\mathbf{x}^T\mathbf{S}_{\lambda_x}(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{S}_{\lambda_x} &= \mathcal{O}(\lambda) + \mathcal{O}(n^{-1}\lambda_x^{-d/2m}\log^2 n) \\ &\approx \mathcal{O}(n^{-\delta}) + \mathcal{O}(n^{-(1-\delta_x d/2m)}\log^2 n) = \mathcal{O}(1). \end{aligned}$$

This proves (f).

3 Web Appendix C: Main asymptotic results

This appendix details the main results of our asymptotic analysis referred to in Section 3 of the paper.

3.1 Asymptotic results for the spatial model

In the model (1) of the paper, spatial correlation is modeled through smoothing of the term f. Without the smoothing penalty, the model is an ordinary linear model in which all effect estimates are unbiased. Therefore, bias in the covariate effect estimate arises as a direct result of smoothing. Rice (1986) showed for dimension $d = 1$ that, while this bias is asymptotically 0 as $n \to \infty$, the rate of convergence may be slow. More specifically, we cannot ensure that the bias converges faster than the standard deviation if the smoothing parameter λ converges at the optimal rate (minimizing the AMSE of the estimated spatial effect). Therefore, the bias can in practice become disproportionately large. Here, we generalize Rice's results and see that the problem of potentially excessive bias in $\widehat{\beta}$ persists in models where the spatial domain has dimension $d \ge 1$. As an aside, we note that, as in the $d = 1$ case, the rate of convergence of the variance of $\widehat{\beta}$, is the same as that in a model with no smoothing penalty.

Theorem 3.1. *Suppose $\lambda \approx n^{-\delta}$ for some $0 < \delta < 1$, $f, f^x \in H^m(\Omega)$ are bounded and $m \ge d$. Then for the partial thin plate spline estimate of β we have that*

(a) $\mathrm{E}(\widehat{\beta}) - \beta = o(n^{-1/2}) + \mathcal{O}(\lambda^{1/2})$,

(b) $n\mathrm{Var}(\widehat{\beta}) \to \sigma^2/\sigma_x^2$ *as* $n \to \infty$.

In particular, $\mathrm{Var}(\widehat{\beta}) = \mathcal{O}(n^{-1})$ *and we need* $\lambda = o(n^{-1})$ *to ensure that the bias converges faster than the standard deviation of* $\widehat{\beta}$.

Proof. Let $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$. Since $\mathrm{E}(\mathbf{y}) = \beta\mathbf{x} + \mathbf{f}$, the expression (3) in the paper shows that

$$\begin{aligned} \mathrm{E}(\widehat{\beta}) - \beta &= \big(\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{x}\big)^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)(\beta\mathbf{x}+\mathbf{f}) - \beta \\ &= \big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{x}\big)^{-1}\big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{f}\big) \\ &= o(n^{-1/2}) + \mathcal{O}(\lambda^{1/2}) \end{aligned}$$

by Lemma 1.4 (a) and (c).

Similarly, since $\mathrm{Var}(\mathbf{y}) = \sigma^2\mathbf{I}$, (3) in the paper shows that

$$\begin{aligned} n\mathrm{Var}(\widehat{\beta}) &= n\sigma^2\big(\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{x}\big)^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{x}\big(\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{x}\big)^{-1} \\ &= \sigma^2\big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{x}\big)^{-1}\big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{x}\big)\big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{x}\big)^{-1} \\ &\to \sigma^2/\sigma_x^2 \quad \text{as } n\to\infty \end{aligned}$$

by Lemma 1.4 (a) and (b). □

Theorem 3.2. *Suppose $\lambda \approx n^{-\delta}$ for some $0 < \delta < 1$, $f, f^x \in H^m(\Omega)$ are bounded and $m \geq d$. Then the average squared bias $B^2(f,\lambda)$ and average variance $V(f,\lambda)$ of the partial thin plate spline estimate of f satisfy*

(a) $B^2(f,\lambda) = n^{-1}\sum_i(\mathrm{E}(\widehat{f}_i) - f(\mathbf{t}_i))^2 = \mathcal{O}(\lambda)$,

(b) $V(f,\lambda) = n^{-1}\sum_i \mathrm{Var}(\widehat{f}_i) = \mathcal{O}(n^{-1}\lambda^{-d/2m})$.

In particular, the optimal rate for λ in terms of minimizing $\mathrm{AMSE}(\widehat{\mathbf{f}})$ is $\lambda = \mathcal{O}(n^{-2m/(2m+d)})$, and when λ converges at this optimal rate, $\mathrm{AMSE}(\widehat{\mathbf{f}}) = \mathcal{O}(n^{-2m/(2m+d)})$.

Proof. Let $\mathbf{f} = (f(\mathbf{t}_1),\ldots,f(\mathbf{t}_n))^T$ and $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^T$ so that $\mathbf{y} = \beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}$.

Since $\widehat{\mathbf{f}} = \mathbf{S}_\lambda(\mathbf{y} - \widehat{\beta}\mathbf{x})$ by (3) in the paper,

$$\mathrm{E}(\widehat{\mathbf{f}}) - \mathbf{f} = -\big(\mathrm{E}(\widehat{\beta}) - \beta\big)\mathbf{S}_\lambda\mathbf{x} - (\mathbf{I}-\mathbf{S}_\lambda)\mathbf{f}.$$

We therefore see that

$$\begin{aligned} B^2(f,\lambda) &= n^{-1}\|\mathrm{E}(\widehat{\mathbf{f}}) - \mathbf{f}\|^2 \\ &\leq n^{-1}\big\|\big(\mathrm{E}(\widehat{\beta}) - \beta\big)\mathbf{S}_\lambda\mathbf{x}\big\|^2 + n^{-1}\|(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{f}\|^2 \\ &= \big(\mathrm{E}(\widehat{\beta}) - \beta\big)^2 n^{-1}\mathbf{x}^T\mathbf{S}_\lambda^2\mathbf{x} + B^2_{\mathrm{tp}}(f,\lambda) \\ &= \big(o(n^{-1}) + \mathcal{O}(\lambda)\big)\mathcal{O}(1) + \mathcal{O}(\lambda) = \mathcal{O}(\lambda) \end{aligned}$$

by Theorem 1(a), Lemma 1.4 (d) and Lemma 1.3. This proves part (a).

For (b), firstly note that

$$\widehat{\mathbf{f}} - \mathrm{E}(\widehat{\mathbf{f}}) = \mathbf{S}_\lambda\boldsymbol{\epsilon} - \big(\widehat{\beta} - \mathrm{E}(\widehat{\beta})\big)\mathbf{S}_\lambda\mathbf{x}.$$

We therefore see that

$$\begin{aligned} V(f,\lambda) &= n^{-1}\mathrm{E}\big(\|\widehat{\mathbf{f}} - \mathrm{E}(\widehat{\mathbf{f}})\|^2\big) \\ &\leq n^{-1}\mathrm{E}(\boldsymbol{\epsilon}^T\mathbf{S}_\lambda^2\boldsymbol{\epsilon}) + \mathrm{E}\big[(\widehat{\beta} - \mathrm{E}(\widehat{\beta}))^2\big]n^{-1}\mathbf{x}^T\mathbf{S}_\lambda^2\mathbf{x} \\ &= n^{-1}\sigma^2\mathrm{Tr}(\mathbf{S}_\lambda^2) + \mathrm{Var}(\widehat{\beta})\mathcal{O}(1) \\ &= \mathcal{O}(n^{-1}\lambda^{-d/2m}) + \mathcal{O}(n^{-1}) = \mathcal{O}(n^{-1}\lambda^{-d/2m}) \end{aligned}$$

by Lemma 1.2, Theorem 1(b) and Lemma 1.4 (d). This proves part (b).

Finally, recall that

$$\mathrm{AMSE}(\widehat{\mathbf{f}}) = B^2(f,\lambda) + V(f,\lambda).$$

From the above, we see that the bias term increases with λ while the variance term decreases with λ so that the optimal rate for minimizing $\mathrm{AMSE}(\widehat{\mathbf{f}})$ is achieved when $\mathcal{O}(\lambda) = \mathcal{O}(n^{-1}\lambda^{-d/2m})$. This leads to an optimal rate of $\lambda = \mathcal{O}(n^{-2m/(2m+d)})$. At this rate for λ, $B^2(f,\lambda)$ and $V(f,\lambda)$ converge at the same rate of $\mathcal{O}(n^{-2m/(2m+d)})$. □

We have therefore proved the following result which shows that we cannot avoid the potential for excessive bias in $\widehat{\beta}$, unless λ converges at a rate slower than the optimal rate of convergence, i.e. unless the smooth term is undersmoothed.

Corollary 3.3. *Suppose $\lambda \approx n^{-\delta}$ for some $0 < \delta < 1$, $f, f^x \in H^m(\Omega)$ are bounded and $m \geq d$. The optimal rate of convergence for λ in terms of minimizing* AMSE($\widehat{\mathbf{f}}$) *is slower than the required rate of $o(n^{-1})$ for ensuring that the bias of $\widehat{\beta}$ converges faster than the standard deviation of the estimate.*

3.2 Asymptotic results for the spatial+ model

In dimension $d = 1$, Chen and Shiau (1991) show that for the model (5) of the paper, the problems identified by Rice disappear. That is, when the parameters λ and λ_x converge at the optimal rate (for minimizing the AMSE of the estimated spatial effect), the bias of the covariate effect estimate $\widehat{\beta}^+$ converges to 0 faster than the standard deviation and, therefore, does not become disproportionately large. We now generalize these results to dimensions $d \geq 1$.

Theorem 3.4. *Suppose $\lambda \approx n^{-\delta}, \lambda_x \approx n^{-\delta_x}$ for some $0 < \delta, \delta_x < 1$, $f, f^x \in H^m(\Omega)$ are bounded and $m \geq d$. Then for the spatial+ estimate of β we have that*

(a) $\mathrm{E}(\widehat{\beta}^+) - \beta = o(n^{-1/2}) + \mathcal{O}((\lambda\lambda_x)^{1/2})$,

(b) $n\mathrm{Var}(\widehat{\beta}^+) \to \sigma^2/\sigma_x^2$ *as* $n \to \infty$.

In particular, $\mathrm{Var}(\widehat{\beta}^+) = \mathcal{O}(n^{-1})$ *and we need* $\lambda\lambda_x = o(n^{-1})$ *to ensure that the bias converges faster than the standard deviation of* $\widehat{\beta}^+$.

Proof. Let

$$\begin{aligned} \mathbf{b} &= (\mathbf{I} - \mathbf{S}_\lambda)(\mathbf{I} - \mathbf{S}_{\lambda_x})\mathbf{x} \\ a_1 &= n^{-1}\mathbf{b}^T(\mathbf{I} - \mathbf{S}_{\lambda_x})\mathbf{x} = n^{-1}\mathbf{x}^T(\mathbf{I} - \mathbf{S}_{\lambda_x})(\mathbf{I} - \mathbf{S}_\lambda)(\mathbf{I} - \mathbf{S}_{\lambda_x})\mathbf{x} \\ a_2 &= n^{-1}\mathbf{b}^T\mathbf{b} = n^{-1}\mathbf{x}^T(\mathbf{I} - \mathbf{S}_{\lambda_x})(\mathbf{I} - \mathbf{S}_\lambda)^2(\mathbf{I} - \mathbf{S}_{\lambda_x})\mathbf{x}. \end{aligned}$$

By Lemma 1.5 (a) and (b), $a_1 \to \sigma_x^2$ and $a_2 \to \sigma_x^2$ as $n \to \infty$. From (7) in the paper we see that

$$\widehat{\beta}^+ = (na_1)^{-1}\mathbf{b}^T\mathbf{y}.$$

Therefore, since $\mathrm{E}(\mathbf{y}) = \beta\mathbf{x} + \mathbf{f}$ where $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))$,

$$\begin{aligned} \mathrm{E}(\widehat{\beta}^+) - \beta &= (na_1)^{-1}\big((\mathbf{b}^T\mathbf{x} - na_1)\beta + \mathbf{b}^T\mathbf{f}\big) \\ &= (na_1)^{-1}\big(\mathbf{b}^T\mathbf{S}_{\lambda_x}\mathbf{x}\beta + \mathbf{b}^T\mathbf{f}\big) \\ &= a_1^{-1}\big(n^{-1}\mathbf{x}^T(\mathbf{I} - \mathbf{S}_{\lambda_x})(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\mathbf{x}\beta + n^{-1}\mathbf{x}^T(\mathbf{I} - \mathbf{S}_{\lambda_x})(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{f}\big) \\ &= o(n^{-1/2}) + \mathcal{O}((\lambda\lambda_x)^{1/2}) \end{aligned}$$

by Lemma 1.5 (d) and (c). This proves part (a).

For part (b), since $\mathrm{Var}(\mathbf{y}) = \sigma^2\mathbf{I}$, we see that

$$\begin{aligned} n\mathrm{Var}(\widehat{\beta}^+) &= n(na_1)^{-2}\mathbf{b}^T(\sigma^2\mathbf{I})\mathbf{b} \\ &= (\sigma^2 a_2)/a_1^2 \\ &\to \sigma^2/\sigma_x^2 \quad \text{as } n \to \infty. \end{aligned}$$

This proves (b). □

Theorem 3.5. *Suppose $\lambda \approx n^{-\delta}, \lambda_x \approx n^{-\delta_x}$ for some $0 < \delta, \delta_x < 1$, $f, f^x \in H^m(\Omega)$ are bounded and $m \geq d$. Then the average squared bias $B_+^2(f, \lambda, \lambda_x)$ and average variance $V_+(f, \lambda, \lambda_x)$ of the spatial+ estimate of f satisfy*

(a) $B_+^2(f, \lambda, \lambda_x) = n^{-1} \sum_i (\mathrm{E}(\widehat{f}_i^+) - f(\mathbf{t}_i))^2 = \mathcal{O}(\lambda) + \mathcal{O}(n^{-1}\lambda_x^{-d/2m} \log^2 n)$,

(b) $V_+(f, \lambda, \lambda_x) = n^{-1} \sum_i \mathrm{Var}(\widehat{f}_i^+) = \mathcal{O}(n^{-1}\lambda^{-d/2m})$.

In particular, the optimal rates for λ and λ_x in terms of minimizing $\mathrm{AMSE}(\widehat{\mathbf{f}}^+)$ *are given by $\lambda = \mathcal{O}(n^{-2m/(2m+d)})$ and $\lambda_x = \mathcal{O}(n^{-2m/(2m+d)}(\log n)^{4m/d})$, assuming the convergence rates for $B_+^2(f, \lambda, \lambda)$ and $V_+(f, \lambda, \lambda_x)$ are equal. When λ and λ_x converge at these rates,* $\mathrm{AMSE}(\widehat{\mathbf{f}}^+) = \mathcal{O}(n^{-2m/(2m+d)})$.

Proof. Let $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ so that $\mathbf{y} = \beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}$. Since by (8) in the paper

$$\widehat{\mathbf{f}}^+ = \mathbf{S}_\lambda \mathbf{y} - \big(\mathbf{S}_\lambda + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\big)\widehat{\beta}^+ \mathbf{x},$$

we have that

$$\mathrm{E}(\widehat{\mathbf{f}}^+) - \mathbf{f} = -(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{f} - (\mathrm{E}(\widehat{\beta}^+) - \beta)\big(\mathbf{S}_\lambda + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\big)\mathbf{x} - \beta(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\mathbf{x}.$$

Since $n^{-1}\big\|(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{f}\big\|^2 = B_{\mathrm{tp}}^2(f, \lambda)$, we therefore see that

$$\begin{aligned}
B_+^2(f, \lambda, \lambda_x) &= n^{-1}\|\mathrm{E}(\widehat{\mathbf{f}}^+) - \mathbf{f}\|^2 \\
&\leq n^{-1}\big\|(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{f}\big\|^2 + \big(\mathrm{E}(\widehat{\beta}^+) - \beta\big)^2 n^{-1}\big\|\big(\mathbf{S}_\lambda + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\big)\mathbf{x}\big\|^2 \\
&\quad + \beta^2 n^{-1}\big\|(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\mathbf{x}\big\|^2 \\
&= \mathcal{O}(\lambda) + \big(o(n^{-1}) + \mathcal{O}(\lambda\lambda_x)\big)\mathcal{O}(1) + \mathcal{O}(\lambda) + \mathcal{O}(n^{-1}\lambda_x^{d/2m} \log^2 n) \\
&= \mathcal{O}(\lambda) + \mathcal{O}(n^{-1}\lambda_x^{-d/2m} \log^2 n)
\end{aligned}$$

by Lemma 1.3, Theorem 3(a) and Lemma 1.4 (f) and (e). This proves part (a).

For (b), note that

$$\widehat{\mathbf{f}}^+ - \mathrm{E}(\widehat{\mathbf{f}}^+) = \mathbf{S}_\lambda \boldsymbol{\epsilon} - \big(\widehat{\beta}^+ - \mathrm{E}(\widehat{\beta}^+)\big)\big(\mathbf{S}_\lambda + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\big)\mathbf{x}.$$

We therefore see that

$$\begin{aligned}
V_+(f, \lambda, \lambda_x) &= n^{-1}\mathrm{E}\big[\|\widehat{\mathbf{f}}^+ - \mathrm{E}(\widehat{\mathbf{f}}^+)\|^2\big] \\
&\leq n^{-1}\mathrm{E}[\boldsymbol{\epsilon}^T \mathbf{S}_\lambda^2 \boldsymbol{\epsilon}] - \mathrm{E}\big[(\widehat{\beta}^+ - \mathrm{E}(\widehat{\beta}^+))^2\big] n^{-1}\big\|\big(\mathbf{S}_\lambda + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{S}_{\lambda_x}\big)\mathbf{x}\big\|^2 \\
&= n^{-1}\sigma^2 \mathrm{Tr}(\mathbf{S}_\lambda^2) + \mathrm{Var}(\widehat{\beta}^+)\mathcal{O}(1) \\
&= \mathcal{O}(n^{-1}\lambda^{-d/2m}) + \mathcal{O}(n^{-1}) = \mathcal{O}(n^{-1}\lambda^{-d/2m})
\end{aligned}$$

by Lemma 1.5 (f), Lemma 1.2 and Theorem 3(b). This proves part (b).

Finally, recall that

$$\mathrm{AMSE}(\widehat{\mathbf{f}}^+) = B_+^2(f, \lambda, \lambda_x) + V_+(f, \lambda, \lambda_x).$$

From the above we see that the bias term increases with λ while the variance term decreases with λ so that the optimal rate for minimizing $\mathrm{AMSE}(\widehat{\mathbf{f}}^+)$ is achieved when $\mathcal{O}(\lambda) = \mathcal{O}(n^{-1}\lambda^{-d/2m})$. This leads to an optimal rate of $\lambda = \mathcal{O}(n^{-2m/(2m+d)})$. Since we have assumed that the convergence rates for $B_+^2(f, \lambda, \lambda_x)$ and $V_+(f, \lambda, \lambda_x)$ are equal, the optimal rate for λ_x is then obtained when $\mathcal{O}(n^{-1}\lambda_x^{-d/2m} \log^2 n) = \mathcal{O}(n^{-2m/(2m+d)})$ which leads to $\mathcal{O}(\lambda_x) = n^{-2m/(2m+d)}(\log n)^{4m/d}$. □

From this we obtain the following result which shows that, unlike $\widehat{\beta}$, the estimate $\widehat{\beta}^+$ does not need undersmoothing to avoid excessive bias.

Corollary 3.6. *Suppose* $\lambda \approx n^{-\delta}, \lambda_x \approx n^{-\delta_x}$ *for some* $0 < \delta, \delta_x < 1$, $f, f^x \in H^m(\Omega)$ *are bounded and* $m \geq d$. *If* λ *and* λ_x *converge at the optimal rates in terms of minimizing* $\mathrm{AMSE}(\widehat{\mathbf{f}}^+)$, *then* $\lambda\lambda_x = o(n^{-1})$. *In particular, the optimal rates for* λ *and* λ_x *ensure that the bias of the spatial+ estimate* $\widehat{\beta}^+$ *converges faster than the standard deviation of the estimate.*

Proof. Theorem 3(b) shows that we need $\mathrm{E}(\widehat{\beta}^+) - \beta = o(n^{-1/2})$ to ensure that the bias converges faster than the standard deviation. Part (a) of the same theorem shows that this required rate can be achieved if $\lambda\lambda_x = o(n^{-1})$. Suppose λ and λ_x converge at their optimal rates from Theorem 4. Then since for any $\epsilon > 0$,

$$n^{-2m/(2m+d)}(\log n)^{4m/d} = o(n^{-2m/(2m+d)+\epsilon}),$$

we have that

$$\lambda\lambda_x = o(n^{-4m/(2m+d)+\epsilon}) = o(n^{-1})$$

if we choose $\epsilon = \frac{2m-d}{2m+d}$. This proves the result. □

4 Web Appendix D: Partial residual estimates

In this appendix we consider, as an aside, the asymptotic behaviour of the partial residual estimates introduced by Denby (1986) and, independently, by Speckman (1988), which are the estimates we obtain using the gSEM approach of Thaden and Kneib (2018). Here we adapt the method used in Sections 3.2 and 3.3 of the paper for estimates in the spatial and spatial+ models to show how the asymptotic results in Chen and Shiau (1991) for the partial residual estimates generalize from the one-dimensional model to dimensions $d \geq 1$. We show that, as is the case for the spatial+ model, the smoothing-induced bias in the covariate effect estimate goes to 0 faster than the standard deviation, i.e. the partial residual estimates also avoid the problem of disproportionate smoothing-induced bias.

For a given value $\lambda > 0$ of the smoothing parameter, the partial residual estimates for the covariate effect β and the unknown smooth effect $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$ in the model (1) of the paper, are defined as

$$\begin{aligned} \widehat{\beta}_{\mathrm{pr}} &= \left(\mathbf{x}^T(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{x}\right)^{-1}\mathbf{x}^T(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{y}, \\ \widehat{\mathbf{f}}_{\mathrm{pr}} &= \mathbf{S}_\lambda(\mathbf{y} - \widehat{\beta}_{\mathrm{pr}}\mathbf{x}) \end{aligned} \tag{4}$$

where $\mathbf{S}_\lambda$ is the smoother matrix. A similar argument to that of Section 2.2 of the paper shows that these estimates are the ones we would obtain in the gSEM if, for simplicity, we used the same smoothing parameter in all regressions. That is, the estimate $\widehat{\beta}_{\mathrm{pr}}$ is the same as the estimated effect in the linear model given by

$$r_i^y = \beta r_i^x + \epsilon_i, \quad \epsilon_i \underset{\text{iid}}{\sim} N(0, \sigma^2)$$

where $\mathbf{r}^x = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{x}$ and $\mathbf{r}^y = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}$ are the residuals after fitting a thin plate spline to $\mathbf{x}$ and $\mathbf{y}$, respectively.

Minor adjustments to the proofs of Theorems 1 and 2 and Corollary 1 for the spatial model estimates lead to the following results. These results show that the asymptotic behaviour of the estimates $\widehat{\beta}_{\mathrm{pr}}$ and $\widehat{\mathbf{f}}_{\mathrm{pr}}$ is the same as that of the corresponding spatial model estimates, except for the rate of convergence of the bias of the covariate effect estimate $\widehat{\beta}_{\mathrm{pr}}$. More specifically, $\mathrm{E}(\widehat{\beta}_{\mathrm{pr}}) - \beta = o(n^{-1/2}) + \mathcal{O}(\lambda)$, whereas $\mathrm{E}(\widehat{\beta}) - \beta = o(n^{-1/2}) + \mathcal{O}(\lambda^{1/2})$ and this difference is enough to ensure that the bias converges faster than the standard deviation when λ converges at the optimal rate (for minimizing the AMSE of the estimated spatial effect).

Theorem 4.1. *Suppose* $\lambda \approx n^{-\delta}$ *for some* $0 < \delta < 1$, $f, f^x \in H^m(\Omega)$ *are bounded and* $m \geq d$. *Then for the partial residual estimate of* β *we have that*

(a) $\mathrm{E}(\widehat{\beta}_{\mathrm{pr}}) - \beta = o(n^{-1/2}) + \mathcal{O}(\lambda)$,

(b) $n\mathrm{Var}(\widehat{\beta}_{\mathrm{pr}}) \to \sigma^2/\sigma_x^2$ *as* $n \to \infty$.

In particular, $\mathrm{Var}(\widehat{\beta}_{\mathrm{pr}}) = \mathcal{O}(n^{-1})$ *and we need* $\lambda = o(n^{-1/2})$ *to ensure that the bias converges faster than the standard deviation of* $\widehat{\beta}_{\mathrm{pr}}$.

Proof. Let $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$. Since $\mathrm{E}(\mathbf{y}) = \beta\mathbf{x} + \mathbf{f}$, the expression (4) shows that

$$\begin{aligned} \mathrm{E}(\widehat{\beta}_{\mathrm{pr}}) - \beta &= \big(\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{x}\big)^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2(\beta\mathbf{x}+\mathbf{f}) - \beta \\ &= \big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{x}\big)^{-1}\big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{f}\big) \\ &= o(n^{-1/2}) + \mathcal{O}(\lambda) \end{aligned}$$

by Lemma 1.4 (b) and Lemma 1.5 (c).

Similarly, since $\mathrm{Var}(\mathbf{y}) = \sigma^2\mathbf{I}$, (4) shows that

$$\begin{aligned} n\mathrm{Var}(\widehat{\beta}_{\mathrm{pr}}) &= n\sigma^2\big(\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{x}\big)^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^4\mathbf{x}\big(\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{x}\big)^{-1} \\ &= \sigma^2\big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{x}\big)^{-1}\big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^4\mathbf{x}\big)\big(n^{-1}\mathbf{x}^T(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{x}\big)^{-1} \\ &\to \sigma^2/\sigma_x^2 \quad \text{as } n \to \infty \end{aligned}$$

by Lemma 1.4 (b) and Lemma 1.5 (b). □

Theorem 4.2. *Suppose* $\lambda \approx n^{-\delta}$ *for some* $0 < \delta < 1$, $f, f^x \in H^m(\Omega)$ *are bounded and* $m \geq d$. *Then the average squared bias* $B^2_{\mathrm{pr}}(f,\lambda)$ *and average variance* $V_{\mathrm{pr}}(f,\lambda)$ *of the partial residual estimate of* f *satisfy*

(a) $B^2_{\mathrm{pr}}(f,\lambda) = n^{-1}\sum_i(\mathrm{E}((\widehat{\mathbf{f}}_{\mathrm{pr}})_i) - f(\mathbf{t}_i))^2 = \mathcal{O}(\lambda)$,

(b) $V_{\mathrm{pr}}(f,\lambda) = n^{-1}\sum_i \mathrm{Var}((\widehat{\mathbf{f}}_{\mathrm{pr}})_i) = \mathcal{O}(n^{-1}\lambda^{-d/2m})$.

In particular, the optimal rate for λ *in terms of minimizing* $\mathrm{AMSE}(\widehat{\mathbf{f}}_{\mathrm{pr}})$ *is* $\lambda = \mathcal{O}(n^{-2m/(2m+d)})$, *and when* λ *converges at this optimal rate,* $\mathrm{AMSE}(\widehat{\mathbf{f}}_{\mathrm{pr}}) = \mathcal{O}(n^{-2m/(2m+d)})$.

Proof. Let $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ so that $\mathbf{y} = \beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}$.

By (4), $\widehat{\mathbf{f}}_{\mathrm{pr}} = \mathbf{S}_\lambda(\mathbf{y} - \widehat{\beta}_{\mathrm{pr}}\mathbf{x})$ has the same format as the corresponding partial thin plate spline estimate, and therefore,

$$\mathrm{E}(\widehat{\mathbf{f}}_{\mathrm{pr}}) - \mathbf{f} = -\big(\mathrm{E}(\widehat{\beta}_{\mathrm{pr}}) - \beta\big)\mathbf{S}_\lambda\mathbf{x} - (\mathbf{I}-\mathbf{S}_\lambda)\mathbf{f}$$

and

$$\widehat{\mathbf{f}}_{\mathrm{pr}} - \mathrm{E}(\widehat{\mathbf{f}}_{\mathrm{pr}}) = \mathbf{S}_\lambda\boldsymbol{\epsilon} - \big(\widehat{\beta}_{\mathrm{pr}} - \mathrm{E}(\widehat{\beta}_{\mathrm{pr}})\big)\mathbf{S}_\lambda\mathbf{x}.$$

as in the proof of Theorem 2. For the derivation of $B^2_{\mathrm{pr}}(f,\lambda)$ and $V_{\mathrm{pr}}(f,\lambda)$, we can therefore apply the same proof where the only adjustment needed is the rate of convergence of the bias $\mathrm{E}(\widehat{\beta}_{\mathrm{pr}}) - \beta$.

$$\begin{aligned} B^2_{\mathrm{pr}}(f,\lambda) &= n^{-1}\|\mathrm{E}(\widehat{\mathbf{f}}_{\mathrm{pr}}) - \mathbf{f}\|^2 \\ &\leq n^{-1}\big\|\big(\mathrm{E}(\widehat{\beta}_{\mathrm{pr}}) - \beta\big)\mathbf{S}_\lambda\mathbf{x}\big\|^2 + n^{-1}\|(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{f}\|^2 \\ &= \big(\mathrm{E}(\widehat{\beta}_{\mathrm{pr}}) - \beta\big)^2 n^{-1}\mathbf{x}^T\mathbf{S}_\lambda^2\mathbf{x} + B^2_{\mathrm{tp}}(f,\lambda) \\ &= \big(o(n^{-1}) + \mathcal{O}(\lambda^2)\big)\mathcal{O}(1) + \mathcal{O}(\lambda) = \mathcal{O}(\lambda) \end{aligned}$$

by Theorem 4.1 (a), Lemma 1.4 (d) and Lemma 1.3. This proves part (a).

$$\begin{aligned} V_{\mathrm{pr}}(f,\lambda) &= n^{-1}\mathrm{E}\big(\|\widehat{\mathbf{f}}_{\mathrm{pr}} - \mathrm{E}(\widehat{\mathbf{f}}_{\mathrm{pr}})\|^2\big) \\ &\leq n^{-1}\mathrm{E}(\boldsymbol{\epsilon}^T\mathbf{S}_\lambda^2\boldsymbol{\epsilon}) + \mathrm{E}\big[\big(\widehat{\beta} - \mathrm{E}(\widehat{\beta}_{\mathrm{pr}})\big)^2\big]n^{-1}\mathbf{x}^T\mathbf{S}_\lambda^2\mathbf{x} \\ &= n^{-1}\sigma^2\mathrm{Tr}(\mathbf{S}_\lambda^2) + \mathrm{Var}(\widehat{\beta}_{\mathrm{pr}})\mathcal{O}(1) \\ &= \mathcal{O}(n^{-1}\lambda^{-d/2m}) + \mathcal{O}(n^{-1}) = \mathcal{O}(n^{-1}\lambda^{-d/2m}) \end{aligned}$$

by Lemma 1.2, Theorem 4.1 (b) and Lemma 1.4 (d). This proves part (b).

The same argument as we used for the partial thin plate spline estimate $\widehat{\mathbf{f}}$ shows that the optimal rate of convergence for minimizing $\mathrm{AMSE}(\widehat{\mathbf{f}}_{\mathrm{pr}})$ is achieved when $\mathcal{O}(\lambda) = \mathcal{O}(n^{-1}\lambda^{-d/2m})$, which leads to $\lambda = \mathcal{O}(n^{-2m/(2m+d)})$ and $\mathrm{AMSE}(\widehat{\mathbf{f}}_{\mathrm{pr}}) = \mathcal{O}(n^{-2m/(2m+d)})$. □

Corollary 4.3. *Suppose $\lambda \approx n^{-\delta}$ for some $0 < \delta < 1$, $f, f^x \in H^m(\Omega)$ are bounded and $m \geq d$. If λ converges at the optimal rate in terms of minimizing* $\mathrm{AMSE}(\widehat{\mathbf{f}}_{\mathrm{pr}})$*, then*

$$\lambda = o(n^{-1/2}).$$

In particular, the optimal rate for λ ensures that the bias of the partial residual estimate $\widehat{\beta}_{\mathrm{pr}}$ converges faster than the standard deviation of the estimate.

5 Web Appendix E: Derivations for unsmoothed models

In this section we consider in more detail the models we compare in Section 4 of the paper when no smoothing penalty is applied (i.e. $\lambda = \lambda_x = 0$) and include some derivations that help explain our simulation results for these models. In the unsmoothed case, the models are ordinary linear models and the estimated covariate effect $\widehat{\beta}$ and fitted values can be found using simple linear algebra. To avoid confusion, rather than using the same notation for the estimated covariate effect, we denote the estimate by $\widehat{\beta}_{\mathrm{null}}$, $\widehat{\beta}$, $\widehat{\beta}_{\mathrm{RSR}}$, $\widehat{\beta}_{\mathrm{gSEM}}$, $\widehat{\beta}^{+}$ in the null, spatial, RSR, gSEM and spatial+ models, respectively.

Recall that, by the data generation process, each replicate of the response data in the simulation is of the form

$$\mathbf{y} = \beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}^y$$

where β and $\mathbf{f} = -\mathbf{z} - \mathbf{z}'$ are the true covariate and spatial effects, respectively, and $\boldsymbol{\epsilon}^y$ is iid noise.

In the null and RSR models the estimated covariate effect is the ordinary least squares estimate, in particular, for any given data replicate, the estimate in these models are identical. More specifically,

$$\begin{aligned} \widehat{\beta}_{\mathrm{null}} = \widehat{\beta}_{\mathrm{RSR}} &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \\ &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T(\beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}^y), \end{aligned}$$

and hence

$$\mathrm{E}(\widehat{\beta}_{\mathrm{null}}) = \mathrm{E}(\widehat{\beta}_{\mathrm{RSR}}) = \beta + \mathrm{E}((\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{f}).$$

So the bias in the null and RSR models is given by $\mathrm{E}((\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{f})$ and is, therefore, directly related to the correlation between the covariate $\mathbf{x}$ and the true unmeasured spatial effect $\mathbf{f}$. Since in our simulations the correlation is negative, the bias in our results is therefore negative. While the covariate effect estimates in the null and RSR models agree, the fitted values differ as the larger model matrix in RSR explains a part of $\mathbf{y}$ that is treated as random noise in the null model. In fact, the column space of the model matrix of the RSR model is the same as that of the spatial model and, therefore, (when $\lambda = 0$) the fitted values in these models agree (i.e. for any given data replicate, the fitted values will be identical).

When no smoothing penalty is applied, the spatial model, the gSEM and the spatial+ model are essentially the same, i.e. for any given data replicate they have the same fitted values and the same unbiased estimate for the covariate effect. The spatial model is an ordinary linear model where the columns in the model matrix are the covariate $\mathbf{x}$ and the spatial basis vectors $\mathbf{B}_{\mathrm{sp}}$. The spatial+ model is a reparametrization of the spatial model where the column $\mathbf{x}$ in the model matrix is replaced by the spatial residuals $\mathbf{r}^x = \mathbf{x} - \widehat{\mathbf{f}}^x$ (where $\widehat{\mathbf{f}}^x$ are the fitted values of a spatial thin plate spline fitted to $\mathbf{x}$). This does not change the overall column space as the difference $\widehat{\mathbf{f}}^x$ lies in the column space of $\mathbf{B}_{\mathrm{sp}}$. By the data generation process,

$$\begin{aligned} \mathbf{y} &= \beta\mathbf{x} + \mathbf{f} + \boldsymbol{\epsilon}^y \\ &= \beta\mathbf{r}^x + \beta\widehat{\mathbf{f}}^x - \mathbf{z} - \mathbf{z}' + \boldsymbol{\epsilon}^y, \end{aligned}$$

with $\beta\widehat{\mathbf{f}}^x - \mathbf{z} - \mathbf{z}'$ in the column space of $\mathbf{B}_{\text{sp}}$ and, therefore, the true effect of $\mathbf{r}^x$ is the same as that of $\mathbf{x}$. In fact, since $\mathbf{r}^x$ is orthogonal to the spatial basis vectors, the estimated effect $\widehat{\beta}^+$ in the spatial+ model (14) of the paper (and therefore $\widehat{\beta}$ in the spatial model (10) of the paper) is obtained as

$$\begin{aligned}\widehat{\beta} = \widehat{\beta}^+ &= (\mathbf{r}^{xT}\mathbf{r}^x)^{-1}\mathbf{r}^{xT}\mathbf{y} \\ &= \beta + (\mathbf{r}^{xT}\mathbf{r}^x)^{-1}\mathbf{r}^{xT}\boldsymbol{\epsilon}^y.\end{aligned}$$

Similarly, for the unsmoothed gSEM, since

$$\mathbf{r}^y = \mathbf{y} - \widehat{\mathbf{f}}^y = \beta\mathbf{r}^x + \beta\widehat{\mathbf{f}}^x - \mathbf{z} - \mathbf{z}' - \widehat{\mathbf{f}}^y + \boldsymbol{\epsilon}^y,$$

with $\beta\widehat{\mathbf{f}}^x - \mathbf{z} - \mathbf{z}' - \widehat{\mathbf{f}}^y$ in the column space of $\mathbf{B}_{\text{sp}}$, the estimated effect of $\mathbf{r}^x$ in the gSEM model (13) of the paper is given by

$$\begin{aligned}\widehat{\beta}_{\text{gSEM}} &= \left(\mathbf{r}^{xT}\mathbf{r}^x\right)^{-1}\mathbf{r}^{xT}\mathbf{r}^y \\ &= \beta + (\mathbf{r}^{xT}\mathbf{r}^x)^{-1}\mathbf{r}^{xT}\boldsymbol{\epsilon}^y.\end{aligned}$$

This shows that $\widehat{\beta} = \widehat{\beta}^+ = \widehat{\beta}_{\text{gSEM}}$. Note that, since $\mathbf{r}^x$ and $\boldsymbol{\epsilon}^y$ are independent

$$\text{E}(\widehat{\beta}) = \text{E}(\widehat{\beta}^+) = \text{E}(\widehat{\beta}_{\text{gSEM}}) = \beta,$$

i.e. the estimated covariate effect is unbiased.

6 Web Appendix F: Additional simulation results

6.1 Mis-specified model

In the simulations of Section 4 of the paper, the data was generated in such a way that the true spatial dependence was that of a thin plate spline. This was ensured by replacing the spatial fields $\mathbf{z}$ and $\mathbf{z}'$ by the fitted values of a thin plate spline model fitted to them. However, in practice, this assumption may not hold and we therefore repeated the simulations for data generated in the same way but where, instead of fitting a thin plate spline model to $\mathbf{z}$ and $\mathbf{z}'$, we used Gaussian process smooths. More specifically, $\mathbf{z}$ has an exponential covariance structure with range parameter 5 and $\mathbf{z}'$ a spherical covariance structure with range parameter 1. Figure 1 shows the results of these simulations. We see that the results are very similar to the simulation results in the paper.

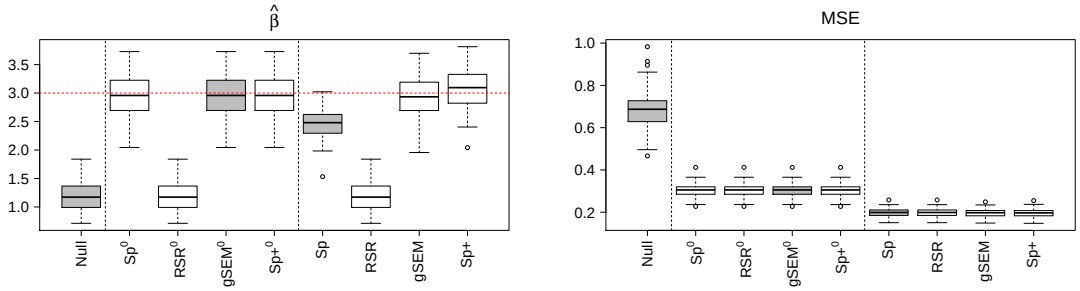


Figure 1: Results of simulations where the true spatial fields $\mathbf{z}$ and $\mathbf{z}'$ are Gaussian process smooths.

6.2 Moderate sample size

In order to investigate the behaviour at moderate sample sizes, we repeated the simulations of Section 4 of the paper for sample sizes $n = 300$, $n = 150$ and $n = 50$ (with spatial basis sizes $k_{\text{sp}} = 100$, $k_{\text{sp}} = 100$ and $k_{\text{sp}} = 30$, respectively). The results shown in Figure 2.

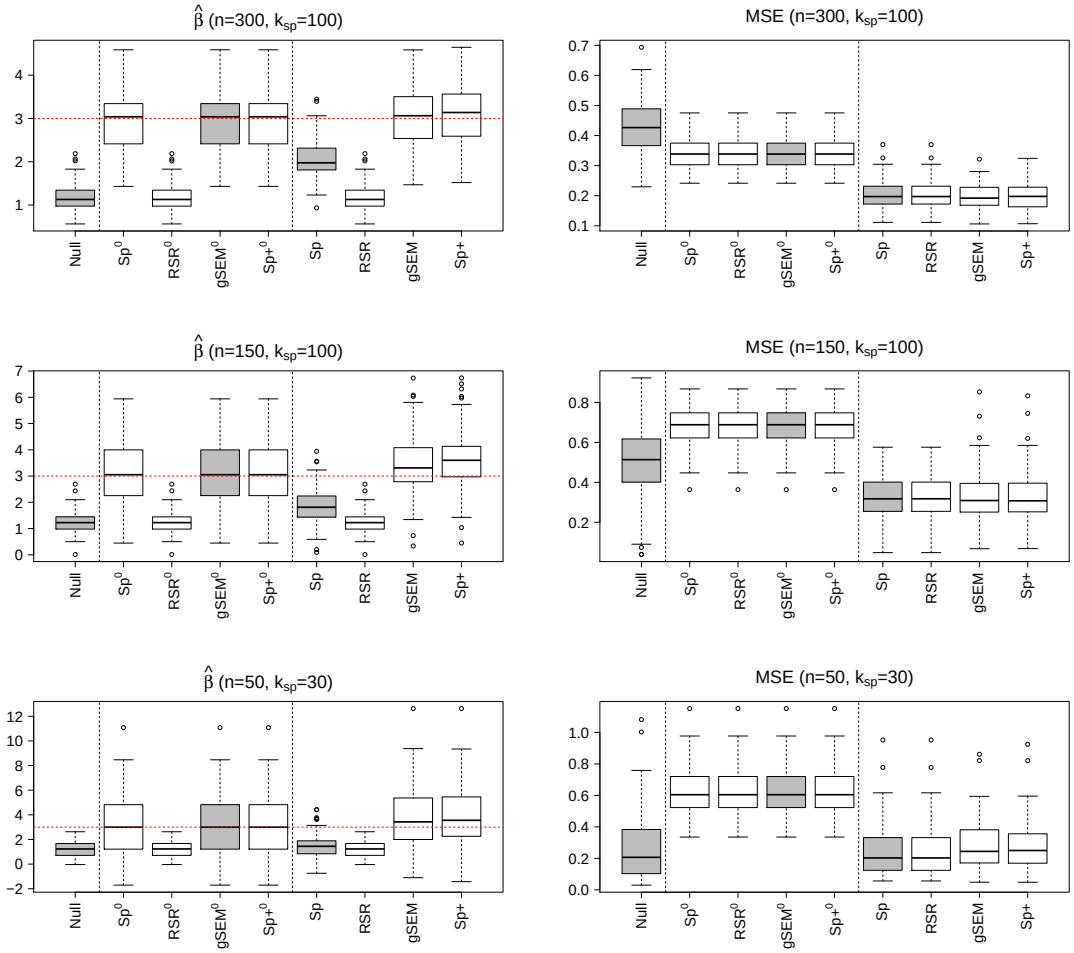


Figure 2: Results of simulations with smaller sample size n.

7 Web Appendix G: Non-Gaussian response data

A distribution is in the exponential family of distributions if its probability density function p can be written in the form

$$p(y) = \exp\left[\{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)\right]$$

where θ and ϕ are parameters of the distribution and a, b and c are functions. This family includes a large number of commonly used distributions in applied statistics, e.g. Gaussian, Poisson, gamma and binomial.

7.1 Spatial model

Suppose we have response data $\mathbf{y} = (y_1, \ldots, y_n)^T$ where each y_i is assumed to be a random variable whose distribution is from the exponential family with $\mathrm{E}(y_i) = \mu_i$, and suppose $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{t}_1, \ldots, \mathbf{t}_n$ are covariate observations and spatial locations as before. A generalized version of (1) in the paper can then be formulated as

$$g(\mu_i) = \beta x_i + f(\mathbf{t}_i) \tag{5}$$

where β is an unknown parameter, f a thin plate spline and $g : \mathbb{R} \to \mathbb{R}$ a link function (i.e. a monotonic smooth function which ensures $g(\mu_i)$ is in the domain of the response variable). The partial thin plate spline estimates of β and $\mathbf{f} = (f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n))^T$ are found using a penalized iterative re-weighted least squares (PIRLS) algorithm. Initializing the algorithm with $\widehat{\mu}_i = y_i$ and $\widehat{\eta}_i = g(\widehat{\mu}_i)$, we define so-called pseudodata as $z_i = g'(\widehat{\mu}_i)(y_i - \widehat{\mu}_i) + \widehat{\eta}_i$ and iterative weights $w_i = 1/(g'(\widehat{\mu}_i)^2 V(\widehat{\mu}_i))$ where $V(\mu_i) = \mathrm{Var}(y_i)\phi = b_i''(\theta)a_i(\phi)/\phi$ is the variance function for the distribution of y_i. Let $\widehat{\beta}$ and $\widehat{\mathbf{f}}$ be the minimizers of

$$\|\sqrt{\mathbf{W}}(\mathbf{z} - \beta\mathbf{x} - \mathbf{f})\|^2 + n\phi\lambda\mathbf{f}^T\mathbf{\Gamma}\mathbf{f} \tag{6}$$

where $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$ is the weights matrix, $\mathbf{z} = (z_1, \ldots, z_n)^T$, and $\lambda > 0$ and $\mathbf{\Gamma}$ are as in (2) of the paper. Now redefining $\widehat{\eta}_i = \widehat{\beta}\mathbf{x} + \widehat{\mathbf{f}}$ and $\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i)$, the algorithm is iterated until convergence and the partial thin plate spline estimates $\widehat{\beta}$ and $\widehat{\mathbf{f}}$ are then the minimizers of (6) in the final iteration. Note that, if no smoothing is applied, $\widehat{\beta}$ and $\widehat{\mathbf{f}}$ are the maximum likelihood estimates in a generalized linear model (GLM), which are asymptotically unbiased.

7.2 Spatial+ model

Starting with the model (5), let $\mathbf{W}$ and $\mathbf{z}$ denote the weights matrix and pseudodata at convergence of the PIRLS algorithm. We then define the corresponding spatial+ model as follows. Let $\widehat{\mathbf{f}}^x$ and $\mathbf{r}^x = \mathbf{x} - \widehat{\mathbf{f}}^x = (r_1^x, \ldots, r_n^x)^T$ denote the fitted values and residuals in the weighted version of the thin plate regression (4) of the paper with weights $\mathbf{W}$, i.e. $\widehat{\mathbf{f}}^x$ is the minimizer of

$$\|\sqrt{\mathbf{W}}(\mathbf{x} - \mathbf{f}^x)\|^2 + n\lambda_x\mathbf{f}^{xT}\mathbf{\Gamma}\mathbf{f}^x$$

with smoothing parameter $\lambda_x > 0$ and $\mathbf{\Gamma}$ defined as before. The spatial+ model is then the partial thin plate spline model defined by

$$g(\mu_i) = \beta r_i^x + f^+(\mathbf{t}_i) \tag{7}$$

where β and f^+ are estimated as described in Section 7.1 above. From Section 7.1 we see that the estimates $\widehat{\beta}$ and $\widehat{\mathbf{f}}$ in the spatial model (5) are obtained as the minimizers of (2) in the paper if we replace $\mathbf{y}, \mathbf{x}, \mathbf{f}, \mathbf{\Gamma}$ and λ by $\widetilde{\mathbf{y}} = \sqrt{\mathbf{W}}\mathbf{z}$, $\widetilde{\mathbf{x}} = \sqrt{\mathbf{W}}\mathbf{x}$, $\widetilde{\mathbf{f}} = \sqrt{\mathbf{W}}\mathbf{f}$, $\widetilde{\mathbf{\Gamma}} = \sqrt{\mathbf{W}}^{-1}\mathbf{\Gamma}\sqrt{\mathbf{W}}^{-1}$ and $\widetilde{\lambda} = \phi\lambda$. Thus, at convergence of the PIRLS algorithm, estimation corresponds to that of a Gaussian model for which the model matrix has columns $\widetilde{\mathbf{x}}$ and $\sqrt{\mathbf{W}}\mathbf{B}_{\mathrm{sp}}$. From our comment at the beginning of Section 4.4 of the paper, the decorrelation trick that we used in Section 2.2 of the paper would therefore work if we replace $\widetilde{\mathbf{x}}$ by $\widetilde{\mathbf{r}}$, obtained from a decomposition $\widetilde{\mathbf{x}} = \widetilde{\mathbf{v}} + \widetilde{\mathbf{r}}$ in which $\widetilde{\mathbf{v}}$ is in the column space of $\sqrt{\mathbf{W}}\mathbf{B}_{\mathrm{sp}}$ and $\widetilde{\mathbf{r}}$ is broadly orthogonal to the columns of $\sqrt{\mathbf{W}}\mathbf{B}_{\mathrm{sp}}$. By the properties of weighted thin plate spline regressions, $\sqrt{\mathbf{W}}\mathbf{r}^x$ is broadly orthogonal to $\sqrt{\mathbf{W}}\mathbf{B}_{\mathrm{sp}}$. Therefore, letting $\widetilde{\mathbf{v}} = \sqrt{\mathbf{W}}\widehat{\mathbf{f}}^x$ and $\widetilde{\mathbf{r}} = \sqrt{\mathbf{W}}\mathbf{r}^x$, the required decorrelation is achieved. Finally, replacing $\widetilde{\mathbf{x}}$ by $\widetilde{\mathbf{r}}$ is equivalent to replacing $\mathbf{x}$ by $\mathbf{r}^x$ in the spatial model, leading to the model (7).

7.3 RSR

Recall that in the Gaussian version of RSR, correlation between the covariate and spatial effect estimates is eliminated by restricting the spatial effect to the orthogonal complement of $\mathbf{x}$. In Section 7.2 we saw that estimation in the generalized version of the spatial model (7) corresponds to that of a Gaussian model in which the model matrix has columns $\widetilde{\mathbf{x}} = \sqrt{\mathbf{W}}\mathbf{x}$ and $\sqrt{\mathbf{W}}\mathbf{B}_{\mathrm{sp}}$ with $\mathbf{W}$ the weights matrix at convergence of

the PIRLS algorithm. We can therefore define the generalized RSR model to be the same as the generalized spatial model but with the spatial basis vectors $\mathbf{B}_{\text{sp}}$ in the model matrix replaced by

$$\widetilde{\mathbf{B}}_{\text{sp}} = (\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{W}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{W})\mathbf{B}_{\text{sp}}.$$

Then, by construction, the generalized RSR model corresponds to a Gaussian model for which the columns $\widetilde{\mathbf{x}} = \sqrt{\mathbf{W}}\mathbf{x}$ and $\sqrt{\mathbf{W}}\widetilde{\mathbf{B}}_{\text{sp}}$ are orthogonal:

$$\widetilde{\mathbf{x}}^T\sqrt{\mathbf{W}}\widetilde{\mathbf{B}}_{\text{sp}} = \mathbf{x}^T\mathbf{W}\widetilde{\mathbf{B}}_{\text{sp}} = \mathbf{0}.$$

References

Chen, H. and Shiau, J.-J. H. (1991). A two-stage spline smoothing method for partially linear models. *Journal of Statistical Planning and Inference* **27,** 187–201.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics & probability letters* **4,** 203–208.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)* **50,** 413–436.

Thaden, H. and Kneib, T. (2018). Structural equation models for dealing with spatial confounding. *The American Statistician* **72,** 239–252.

Utreras, F. I. (1988). Convergence rates for multivariate smoothing spline functions. *Journal of approximation theory* **52,** 1–27.

Closing remarks for Paper 1

The spatial confounding literature tends to focus on the case where the covariate of interest is fully determined by spatial location. In this situation, separating covariate effects from spatial effects is difficult and, while much research has gone into understanding the drivers behind the resulting bias, methodology for removing such bias is yet to be established in this case. The results in this paper are novel as they provide a theoretically-backed and easily implementable method for dealing with spatial confounding bias in a situation often encountered in practice, namely, when the covariate is spatially dependent but not fully determined by spatial location.

This not fully spatial covariate structure is also assumed for the gSEM method proposed by Thaden and Kneib [2018]. The covariate effect estimate in gSEM is obtained through a regression involving only the residuals after all spatial information is removed from both the covariate and the response data. Using simulations, Thaden and Kneib show that this estimate appears to be broadly unbiased. As an aside (included in an appendix of the paper), we investigated the asymptotic behaviour of the estimates in gSEM in a similar way to our analysis of the spatial and spatial+ models. This confirmed that the gSEM estimate has negligible bias and, thus, our analysis provides the theoretical explanation for the behaviour observed in simulations. However, the model seems less intuitive than spatial+ and, as the response variable is different to that of the null and spatial models, standard model selection criteria can no longer be used for comparisons. The spatial+ model also has the advantage that it generalises to non-Gaussian response distributions.

Chapter 4

Paper 2 - Areal models for spatially coherent trend detection: the case of British peak river flows

Introduction to Paper 2

Data collected at different spatial locations are common in many areas of applied statistics and spatial models are becoming an increasingly popular statistical tool. This paper provides an example where a spatial model is used for data pooling, which allows information from different data locations to be shared, improving the inference at each location. Here, we model river flow data collected at gauging stations across Great Britain. The frequency of recent flood events as well as climate models would suggest that flood risk has been increasing over time. However, as the data series at each gauging station is typically quite short and the year-on-year variation high, it is difficult to extract a significant statistical trend signal. Therefore, a time trend had not previously been verified from the flow data. In our paper, we define a spatial model that allows information about such trends to be shared between gauging stations such that if nearby stations have a similar trend, the evidence for a time trend is strengthened. Using this approach, we are able to detect a significant positive time trend in the annual maximum peak flow series, confirming, for the first time, that flood risk has been increasing over time. Moreover, the model identifies the geographical regions with the strongest trend.

Initially, we consider the data at each gauging station separately. Using the approach of Prosdocimi et al. [2014], we fit a linear model with the log annual maximum peak flow as the response variable and time as the only explanatory variable. More precisely, if for a given gauging station we have measurements $Q_1, \ldots, Q_n$ of the annual maximum peak flow at time points $t_1, \ldots, t_n$ (measured in water years), then we fit the model

$$y_i = \log(Q_i) = \alpha + \beta t_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$ is iid noise and α, β, σ are unknown parameters. Using standard linear model theory, we can test the significance of time as an explanatory variable by using a two-sided hypothesis test with null hypothesis $H_0 : \beta = 0$ (i.e. the null hypothesis assumes there is no time trend). Under the null hypothesis, the test statistic is given by

$$T = \frac{\hat{\beta}}{\mathrm{sd}(\hat{\beta})}$$

and has a t_{n-2}-distribution. Here $\hat{\beta}$ denotes the standard linear model estimator for the parameter β and $\mathrm{sd}(\hat{\beta})$ the unbiased estimator for the standard deviation of $\hat{\beta}$. We have only included gauging stations for which the sample size n is relatively large (at least 20) which means that we can assume T is approximately standard normal. If the p-value (i.e. twice the probability that a standard normal variable takes values greater than the observed value of $|T|$) is less than the chosen significance level of 10%, then the null hypothesis of no trend can be rejected, i.e. there is a significant time trend in the annual maximum flow.

Performing this test at each of the 640 gauging stations included in the study we find that, for the vast majority of stations (71%), the test statistic is not significant. However, the fact that the test does not detect a trend does not necessary mean that a trend does not exist. In fact, as pointed out in Prosdocimi et al. [2014], at most gauging stations, the statistical power of the test is quite low which means that it can only detect signals that are relatively strong. The observed value of the test statistic T at a gauging station can be seen as a summary statistic that captures both the strength of the evidence for a time trend at that station as well the direction of the trend. That is, the larger the value of $|T|$, the more evidence there is for a trend and the sign of T is the same as the sign of the estimated slope $\hat{\beta}$. Strong evidence of a trend, i.e. a large value of $|T|$, is obtained when $|\hat{\beta}|$ is large (i.e. the size of the estimated trend is large) or when $\text{sd}(\hat{\beta})$ is small (i.e. when the uncertainty of the trend estimate is small). Since the flow records are typically quite short (on average around 30-40 years) and the year on year variability high, $\text{sd}(\hat{\beta})$ tends to be large, and this means that, even when a non-zero trend exists, $|T|$ may not be large enough for the trend signal to be considered significant.

However, in the above approach, each hypothesis test only uses information from one gauging station. In this paper we implement a method, known as partial pooling, which allows information on the trend signal to be shared across different stations. Partial pooling is a relatively well-established method for information sharing, details of which can be found, for example, in Gelman and Hill [2006]. Rather than performing a test at each station separately, we investigate the evidence for a time trend across Great Britain by modelling the test statistic T in a spatial model, specifically an areal model based on hydrometric areas (HAs). The UK is partitioned into 107 HAs for the purposes of river flow measurements and hydrometric data collection, and our study includes stations from 90 of these HAs. As stations within the same HA are expected to experience similar climate and have similar geophysical properties, we expect the trend signal for stations from the same HA to be more similar than for stations from different HAs. We define the spatial model as

$$T_i = \mu + h_j + \eta_i$$

where T_i is the test statistic at station i, $h_j \sim N(0, \sigma_H^2)$ is an iid random effect for the HA j to which station i belongs, and $\eta_i \sim N(0, \sigma_T^2)$ is the station-specific random error. The parameters μ, σ_H, and σ_T are estimated from the data. Thus, the model assumes that within HA j, the test statistic varies around the mean $\mu + h_j$ with standard deviation σ_T and each h_j varies around the intercept μ with standard deviation σ_H.

Using a Bayesian implementation of this model we obtain an estimate for the test statistic in HA j as

$$\widehat{T_j} = \hat{\mu} + \hat{h}_j$$

where $\hat{\mu}$ and $\hat{h}_j$ are the posterior means of the intercept parameter μ and the random effect h_j, respectively. The estimate $\hat{\mu}$ is approximately "the mean of the HA means" of the test statistic, i.e. if $T_{j_1}, \ldots, T_{j_{N_j}}$ denote the test statistics in the HA j then

$$\overline{T}_j = \tfrac{1}{N_j} \sum_{i=1}^{N_j} T_{j_i}$$

is the corresponding HA mean and

$$\hat{\mu} \approx \tfrac{1}{M} \sum_{j=1}^{M} \overline{T}_j$$

where $M = 90$ is the total number of HAs. Thus, $\hat{\mu}$ is a measure of the overall mean trend signal across all HAs. It can be shown that

$$\widehat{T}_j \approx \frac{\sigma_H^{-2}}{\sigma_T^{-2} N_j + \sigma_H^{-2}} \hat{\mu} + \frac{\sigma_T^{-2} N_j}{\sigma_T^{-2} N_j + \sigma_H^{-2}} \overline{T}_j,$$

i.e. the estimated trend signal in HA j is approximately a weighted average between the overall signal $\hat{\mu}$ and the HA mean $\overline{T}_j$ where the weights depend on the number of stations N_j in the HA. If N_j is large, the HA mean $\overline{T}_j$ is likely to be a relatively reliable estimate and is given a large weight. However, if N_j is small, $\overline{T}_j$ is more likely to be spuriously too high or too low and more weight is put on the overall signal $\hat{\mu}$.

Fitting this model to the data we see that the 90% credible interval for the parameter μ is (0.64, 0.91), indicating an overall tendency for increasing time trends across all HAs. For 54 of the 90 HAs, the entire 90% credible interval for the mean test statistic is positive, and for no HA is the 90% credible interval entirely negative. Thus, there is strong evidence for a positive time trend in peak river flows across most of Great Britain. The model identifies northern England, parts of Scotland, and Wales as the areas with the strongest trend signal, and Southern and Central England as those with the weakest signal.

The paper also includes the results when different temporal subsets of the data are used as well as the results of fitting the model to test statistics obtained through robust regression methods. We see that the overall conclusion of a positive time trend still holds in these cases.

Paper 2

Statement of Authorship

This declaration concerns the article entitled:			
Areal models for spatially coherent trend detection: the case of British peak river flows			
Publication status (tick one) **Draft manuscript** ☐ **Submitted** ☐ **In review** ☐ **Accepted** ☐ **Published** ☒			
Publication details (reference)	Geophysical Research Letters		
Copyright status (tick the appropriate statement) I hold the copyright for this material ☒ Copyright is retained by the publisher, but I have been given permission to replicate the material here ☐			
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	This article was predominantly a joint project between the first author and the candidate with mainly advisory contributions from the remaining co-authors. The candidate contributed to: Formulation of ideas and design/implementation of methodology: - The idea of using spatial modelling for the test statistic in this application had been formulated before the candidate joined the project. However, the candidate prepared the data for implementation, implemented the methodology in practice and contributed with analysis and ideas for implementation, improvements and presentation. 50% Presentation of data in journal format: - The first author prepared the journal format of the article based on the previous work of the candidate. The candidate also contributed with wording and additional computations used for responding to questions posed by the referees. 20%		
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.		
Signed		**Date**	13/2/2021

Areal Models for Spatially Coherent Trend Detection: The Case of British Peak River Flows

Ilaria Prosdocimi[1,2], Emiko Dupont[2], Nicole H. Augustin[2], Thomas R. Kjeldsen[3], Dan P. Simpson[4], and Theresa R. Smith[2]

[1]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy, [2]Department of Mathematical Sciences, University of Bath, Bath, UK, [3]Department of Architecture and Civil Engineering, University of Bath, Bath, UK, [4]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Abstract With increasing concerns on the impacts of climate change, there is wide interest in understanding whether hydrometric and environmental series display any sort of trend. Many studies however, focus on the analysis of highly variable individual series at each measuring location. We propose a novel and straightforward approach to trend detection, modelling the test statistic for trend at each location via an areal model in which the information across measuring locations is pooled together. We exemplify the method with a detailed study of change in high flows in Great Britain. Using areal models, we detect a statistically relevant signal for a positive trend across Great Britain in the recent decades. This evidence is also found when different temporal subsets of the records are analysed. Further, the model identifies areas where the increase has been higher or lower than average, thus providing a way to prioritise intervention.

Plain Language Summary With growing concerns over the potential impacts of climate change, many studies are investigating whether river extremes, such as floods, are changing. Studies based on climate change projections indicate that changes might be expected in several parts of the world, including Great Britain where floods are predicted to increase. However, studies investigating measured river flow records have mostly found inconclusive evidence of change. This does not mean that change is not happening, but finding the evidence of this change is difficult because flow records are short and very variable. In this study we suggest that river flow measuring stations on the same river will experience similar changes since they are affected by the same climate. We therefore propose to use advanced statistical models, which combine information from nearby stations and apply these model to high flows measurements in Great Britain. The analysis of data from closely located measuring stations demonstrates that flows have generally become bigger in Great Britain recently. The methods proposed in the manuscript could be easily applied to other type of data routinely measured and which might have been changing over time as a result of climate change or other drivers.

1. Introduction

River flooding is a major natural hazard that threatens the well-being of communities and can have extremely high costs: The global annual average loss from river flooding is estimated to be USD 104 billion (United Nations Office for Disaster Risk Reduction (UNISDR), 2015), and in the United Kingdom alone, the expected annual flood damages is GBP 560 million (Sayers et al., 2015). There is a widespread interest in understanding how climate change impacts fluvial flood risk (IPCC, 2012) so that appropriate management strategies can be put in place. This interest has resulted in a number of studies investigating projected and observed changes in peak flow magnitude (and/or frequency) at the global (Hirabayashi et al., 2013; Do et al., 2017), continental (Alfieri et al., 2015; Mediero et al., 2015), and national or regional scale (Giuntoli et al., 2015; Slater & Villarini, 2016; Kay et al., 2014; Prosdocimi et al., 2014). The overall picture gives mixed results, with high flows projected to increase and decrease in different areas of the world under representative concentration pathway RCP8.5 (Dankers et al., 2014), while for the U.K. national scale investigations based on the UKCP09 projections (Murphy et al., 2009) under a range of emission scenarios (Kay et al., 2014; 2014) indicate an overall increase in high flows in the last decades of the 21st century. In contrast, studies based on gauged historical data give a more faceted picture, in the United Kingdom as well as in other parts

of the world (Archfield et al., 2016; Hall et al., 2014; Hannaford, 2015), with no clear detectable changes in the behavior of high flows.

Failure to detect a clear time trend signal in gauged peak flows (or other environmental variables) does not necessarily mean that an overall trend does not exist: The absence of evidence for change does not give evidence for the absence of change. Most statistical approaches used for trend detection would need very long records to perform optimally (Svensson et al., 2006), and such long records are sparse in Britain (see Figures S1 and S2 in the supporting information) and generally across the world. In particular, tests applied to short time series have low statistical power; that is, they are not able to detect signals of change even when these are present in the data (Prosdocimi et al., 2014; Vogel et al., 2013). To overcome this lack of power, we develop an areal model that pools information across stations in the same geographical region to enhance the shared trend signal. Areal models can be viewed as multilevel or hierarchical models (see Gelman et al., 2013, Verbeke & Molenberghs, 2009), which are routinely used in life sciences and social sciences to obtain a clearer estimation of the phenomena under study by pooling together the information across several observations (see, e.g., Gelman & Hill, 2012). By pooling together the information of nearby stations, the signal for the evidence of change, and in particular of an increase in flow magnitudes, is enhanced and becomes very clear.

2. Data

We use the annual maxima of the instantaneous (15-min) gauged peak flow recorded at 640 stations in Great Britain (GB) made available by the National River Flow Archive (2018). This is a subset of the national *Peak Flow Dataset*, which is maintained by the National RiverFlow Archive (NRFA) and is the successor of HiFlows-UK, the reference data set used in the United Kingdom to carry out flood estimation studies (Environment Agency, 2012; Lamb et al., 2009). Annual maxima are selected as the highest flow value registered in any given water year, which in the United Kingdom runs from 1 October to 31 September. In this study we used flow values for all the years of station records deemed to have reliable rating curves up to, at least, bank full flow. This ensures that the data series that the measuring authorities deem to be of the highest quality and reliable throughout the recording period are included in the study. To ensure that the results can be indicative of the impacts of (anthropogenic) climate change, only records that end in a year subsequent to the water year 2000 and that refer to catchments with low levels of urban land-cover are included. Finally, only stations with more than 20 years of data are retained in the study. This results in the inclusion of a total of 640 stations with a median length of 47 years: See Text S1 for additional information on the spatiotemporal coverage of the records used in this study.

For practical reasons, river flow measurement and hydrometric data collection in the United Kingdom are organized on a catchment or basin basis, rather than according to the administrative boundaries. Therefore, the country has been divided into 107 hydrometric areas (HA; National River Flow Archive, 2014), which consist in integral river catchments having one or more outlets to the sea or tidal estuary. Of the 107 British HAs, 97 are located in mainland Britain and stations with high-quality annual maxima records are available in 90 of those. Each station is located in a specific HA, and these are defined based on river systems, which typically experience similar climate and weather (see Text S3 for an exploration of the climatology of the HAs), with some of the catchments within each HA possibly nested within each other (and therefore not independent from each other). HAs are based on geophysical properties of river basins and were designed to facilitate an integrated approach to the collection of hydro-meteorological data: Their definition is independent of the study of trends in river flow and as such is an objective way to separate stations into groups that can be expected to behave similarly. We will therefore use the hydrometric areas in the spatial model outlined in the next section. Figure S2 shows how the different hydrometric areas span across the countries in GB.

3. Methods

For each station in the study a simple regression is performed on the log-transformed river flow with time as a covariate, as in Vogel et al. (2011) and Prosdocimi et al. (2014). For each station i, the value of the test statistic for the significance of time T_i is derived. Time here is used as a proxy for anthropogenic climate

change, and the test statistic T_i is a standardized summary of the evidence in favor of a time trend, so of a change, at each station i (see Text S2 for more discussion on the derivation of the test statistic). Stations are located in one HAs only, with each HA typically experiencing similar climate and weather (see Text S3). It is therefore conceivable that similar changes occur at different locations within each HA, so that the test statistic value of stations within each HA should be similar in sign and magnitude and can be pooled together to give a clearer indication for the potential of change in the specific HA and across GB.

An areal model for the test statistic is therefore constructed so that the value of the test statistic at each station is modeled as the random variation around the sum of the average value μ and an areal component h_j, which can take different values for each HA j. This is written as (see, among others, Lawson, 2013)

$$T_i = \mu + h_{j(i)} + \eta_i, \tag{1}$$

where μ is the mean signal for trend across HAs, $h_{j(i)}$ is a parameter taking specific value for the hydrometric area j to which the station i belongs, and $\eta_i \sim N(0, \sigma_T^2)$ is the station-specific random error. This model implies that the test statistic at each station i in a region j is the realization of a random variation around the regional value $\mu + h_j$. It is assumed that the effects h_j for each hydrometric area are independent and identically distributed (iid) with $h_j \sim N(0, \sigma_H^2)$. The h_j's are unknown random quantities that reflect our belief that variability of the test statistic within region j is likely to be smaller than the overall variability of the test statistic. The parameters that need to be estimated from the data are μ, σ_H, and σ_T: This is done in a Bayesian fashion using R-INLA (Rue et al., 2009), which allows for fast approximate estimation of complex models. This means that the posterior distributions of the model parameters given the observed data (i.e., the observed test statistic values) are estimated. Stations within each HA would then have the same estimated posterior distribution for the test statistic in the areal model, an indication of the strength of evidence for a trend in an HA averaged across all stations within the area. From this posterior probability, the evidence for either a positive, negative, or null trend can be derived.

The parameters are estimated by pooling the information from all stations in the network, thereby using the available information in an optimal way. The overall level μ gives an indication of the strength of evidence in favor of a trend across the parts of GB included in this study. More specifically, the posterior estimate of μ is approximately the average of all HA sample averages (where by "HA sample average" we mean the average of the observed test statistics within a given HA). In particular, the pooling in the area-level model means that the posterior estimate of μ is robust to differences in the number of stations per HA. For a given HA, the posterior estimate of the test statistic in this HA is approximately the weighted sum of its HA sample average and the estimated overall trend μ. The weight on the HA sample average increases as the number of stations in the HA increases, meaning the posterior evidence of trend in an HA with many stations is less influenced by pooling than in HAs with sparser data. Details of the estimation theory for partial pooling models such as the areal model presented in equation (1) can be found in Gelman et al. (2012), chapter 12.

A number of approaches to pool information in space have been proposed for the detection of trends in environmental variables (see, e.g., Fischer & Knutti, 2014; Renard et al., 2008), and some of these make use of Bayesian hierarchical models (e.g., in Brady et al., 2019; Renard et al., 2006). The areal model proposed has the advantage of using as the response variable the test statistic, a simple concept that is typically easy to compute, is normalized, and has a well-defined theoretical distribution under the null hypothesis of no change. After choosing a spatial aggregation unit (in this manuscript, the externally predetermined HA), it is straightforward to derive information about the posterior distribution of the average test statistic at each aggregation unit and to identify the areas with high probabilities for the test statistic to be different from 0, that is, an indication of change in the original variable of interest. In this study we propose to use HAs as the spatial aggregation unit, as these have been defined independently for hydrometry purposes and are commonly used in practice to identify river basins and coherent areas for water management purposes. Other aggregations might be used, possibly not based on geographical proximity, but based on, for example, flood-generating mechanism or other similarity measure. Nevertheless, results for different aggregations would be more difficult to visualize on a map, and the interpretation of the results would be less direct since it would not be related to a specific area and river basin.

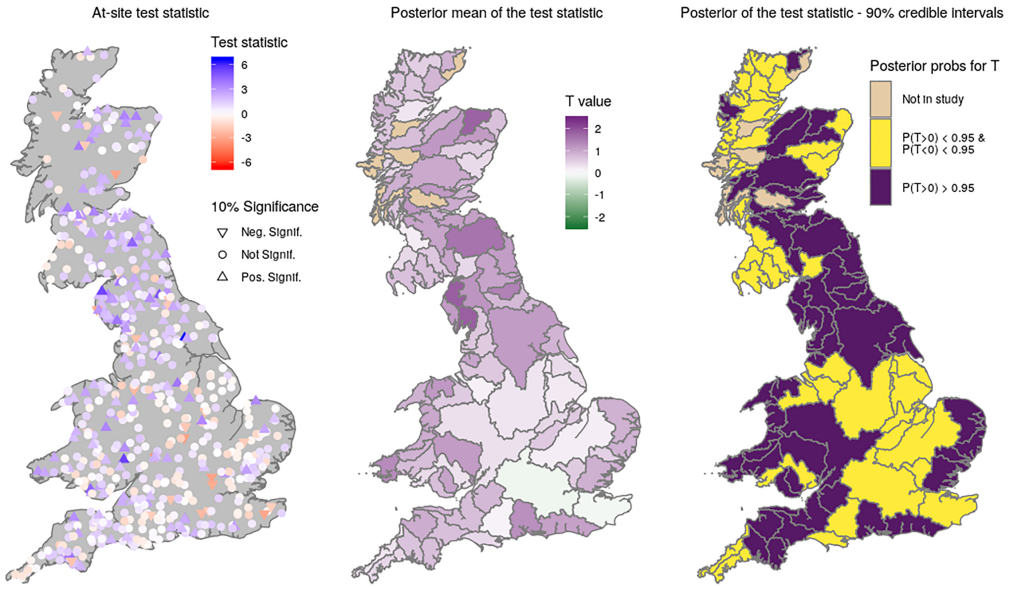


Figure 1. Left panel: at-site test statistic and significance at 10% level for all stations. Central panel: estimated posterior mean derived from the proposed areal model for each area-specific test statistic value. Right panel: summarized information for the 90% credible interval for each area-specific test statistic value.

4. Results

Figure 1 (left panel) exemplifies the ambiguous results typically found when applying a statistical test on a site by site basis to all stations in a river gauging network. The figure shows the values of a test statistic for the time trend derived according to the method outlined in section 3 and further discussed in Text S2.

For a vast majority of stations (71%), the test statistic is not significant at the 10% significance level, indicating that the null hypothesis of no change (i.e., no trend) in time cannot be rejected. As discussed in Prosdocimi et al. (2014), this might be connected to the low statistical power of the test applied to short time series. For 4% of stations a significant negative trend is found, while positive significant trends are found in 25% of stations. There is therefore an indication that positive trends are more frequent than negative trends, and there appears to be some spatial clustering of positive trends in northwestern England and parts of Scotland. The tendency of the test statistic of all stations to be positive rather than negative is also evident in the general distribution of the test statistics, which is shown in Figure S3.

The central and right panel of Figure 1 summarize key results of the areal model fit, highlighting a clear positive trend signal when regional information is pooled together (estimates for the variance components are presented in Table S1 and Text S5). The map in the middle panel shows the mean value of the estimated posterior distribution of the test statistic for each HA: These tend to be positive, with only few areas exhibiting slightly negative values. The 90% credible interval for the overall trend μ is (0.64, 0.91). Thus, there is a tendency for increasing trends across the river flow measuring network in the country. For 54 out of 90 areas, the entire 90% credible interval for the mean test statistic is positive, that is, more than 95% of the posterior distribution of the area-specific test statistic value is larger than 0 (purple HAs in the right panel of Figure 1). For no HA in the country does the 90% credible interval of the marginal posterior distribution of the area-specific test statistic contain negative numbers; this shows that across the river flow measuring network in GB, there is an either null or positive trend. The strongest signal in favor of trend is found in northern England, parts of Scotland, and Wales, and the weakest signal is found in Southern and Central England. This indicates that these areas might need to be given higher, respectively lower, priority for a new flood risk assessment. Some spatially structured variation in the estimated strength of the trend in the different HAs can be noted, even though the model does not specifically enforce this. This might indicate that large-scale climate variability, which operates on a large spatial scale, is a large driver of the changes in high flows. These findings are not dissimilar when robust regression approaches are used in the derivation of the test statistic (see Text S6).

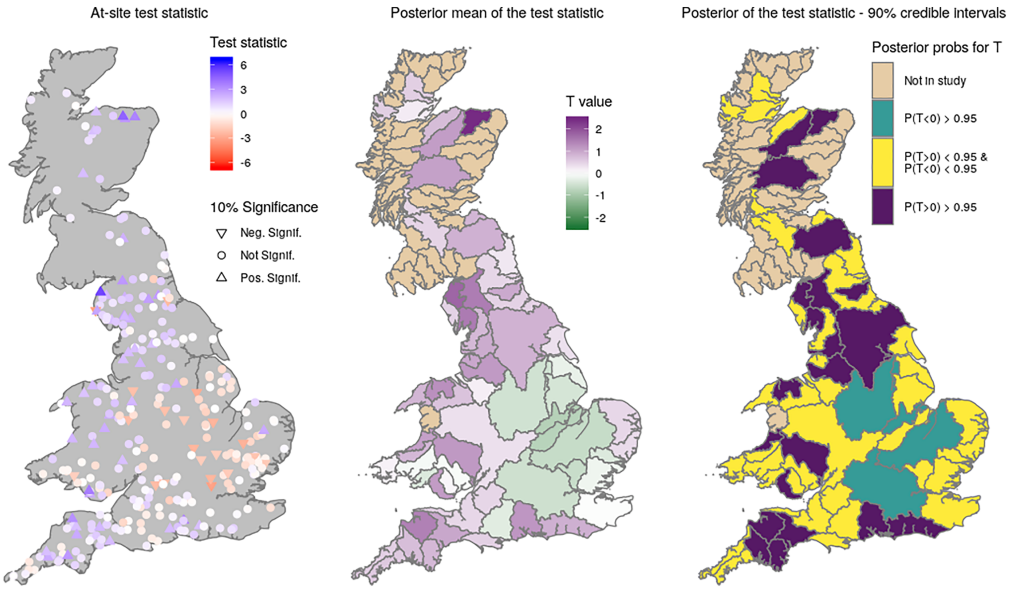


Figure 2. Results for a long common time period analysis (1976–2016). Left panel: at-site test statistic values and significance at 10% significance level for all stations. Central panel: estimated posterior mean derived from the proposed areal model for each area specific test statistic value. Right panel: summarized information for the 90% credible interval for each area specific test statistic value.

The wide range of posterior mean values in the different HAs is possibly the result of very different patterns of change for high flows in different areas of the United Kingdom. This diversity in trend directions has already been highlighted (Hannaford, 2015), but the areal model allows to separate out an island wide effect and the areas that have experienced coherent changes in high flows. Nevertheless, a more HA specific analysis would be needed to identify the possible causes behind the evidence for change (or lack thereof) in any area: Local factors and the response of single catchments to external forcings can have strong impact in the final estimated value of the test statistic for each station in the HA. These local factors are not directly included in the areal model but would need to be taken into account in any assessment of the evidence for a trend within a HA.

The period of record covered by the data can have an influence on the estimated magnitude and sign of the tests, which aim to identify monotonic trends (Hannaford et al., 2013; Svensson et al., 2006), and tests applied to data covering different periods might give contrasting results. As seen in Figures S1 and S2, the flow series available in GB cover different periods of time, with a few very long records and most stations having valid records starting in the 1970s. The overall trend μ and the HA specific signals found in the analysis might therefore be representative of different types of changes, and the strong evidence for trend cannot be directly related to a change in peak flow behavior over a specific period of time. Therefore, we carry out a second analysis that focuses on a subset of stations over a fixed period of time. The analysis uses the 298 stations with complete records between 1976 and 2016 (included), that is, with a total of 41 consecutive years of data. The location of the gauging stations included in the study and the value for the time trend test statistic at each station are shown in Figure 2, together with results of the areal model fitted to the data subset (estimates for the variance components are presented in Table S2 and Text S5). The 90% credible interval for the overall trend signal across the river flow measuring network in GB μ is now found to be (0.31, 0.72): The evidence for trend is not as large as when all records are used, but it is still strong and positive. The posterior mean of the test statistic is found to be negative in 15 out of 65 areas, with the entire 90% credible interval below 0 in 4 of them (the green HAs in the right panel in Figure 2). Changing the time window of the investigation gives a less striking result but still indicates that overall peak flow magnitude is increasing throughout the country.

To further assess the evidence in favor of a changing behavior of peak flows, the subset of stations with exactly 41 years of data was further analyzed taking 10 subsets of 31 consecutive years of data with changing initial year (from 1976 to 1985). The estimated posterior distribution for the overall trend parameter μ in the different subperiods is shown in Figure 3: Across all subperiods the overall trend is generally positive, and for no subperiod does the 90% credible interval contain 0. The lowest posterior mean value (0.23)

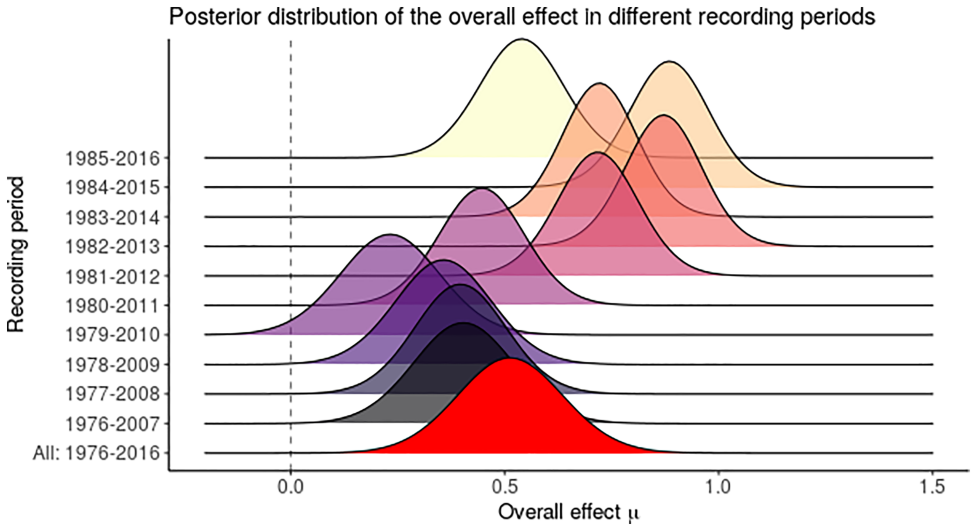


Figure 3. Estimated posterior distributions of μ when using different 31-year-long subsets and the 41-year-long subset in the period 1976–2016.

is found when analyzing the 1979–2010 subperiod, and the highest value (0.88) is found when analyzing the 1984–2015 subperiod. The water year 2010 was characterized by a drought condition (Kendon et al., 2013), while several record-breaking flood events were recorded in 2015 (Barker et al., 2016). Notice also that 1984 was characterized by strong drought conditions (Marsh & Lees, 1985): This might further enhance the strength of the signal for the 1984–2015 period. The difference in the overall effect in the two periods is likely to be a reflection of the general behavior of peak flows in the final and start year of the analysis. In general, the analysis ending in water year 2007 to 2010 indicates an increase in high flows with a smooth decline in time for the overall trend describing the increase. In contrast, the analysis based on records ending in the most recent 6 years have stronger signals in favor of a change with more variability across each subanalysis. This indicates that the overall signal μ increases in each subanalysis, culminating in a very large estimated value μ found when the record-breaking water year 2015 is included in the analysis. This very strong indication for an increase in flood risk is then followed by a much milder signal when the records including the more modest water year 2016 are also included in the analysis. The estimated area-specific posterior mean found for each data subset is shown in Figure S5, with the summary of the credible interval in Figure S6. Regardless of the observation period used in the analysis, there is an indication that peak flow magnitudes are increasing across GB, with a stronger and more persistent signal in the northern part of England and parts of Scotland, while there appear to be less of a concern for changes in high flows in the southeast of England. This finding still holds true when the test statistics included in the areal model are derived from a robust regression model (see Text S6). Even when ensuring that the large records in some series in the latter years are less influential in the estimation of the regression model at each station, a strong evidence for an increase in peak flow is found.

The length of the period for which it is possible to run subanalyses in which a considerable number of stations has a complete record is unfortunately fairly limited and does not allow for more in-depth analyses of the possible large-scale climatic drivers linked with unusually high or low peak flows at a country-wide scale. Climate modes typically evolve slowly in time with persistent periods of positive or negative anomalies, which can impact the behaviors of high flows. For example, modes of the Atlantic Multidecadal Oscillation (AMO) and of the North Atlantic Oscillation (NAO) have been linked to period of elevated high flows in Europe and North America (Hodgkins et al., 2017) and in GB (Hannaford, 2015), thus linking the occurrence of flood-rich periods to multidecadal variability rather than to long-term time trends. Given that in the short time scales for which most flow records are available climate indices have been slowly varying, the detected changes might be a consequence of the dominance of a climatic state rather than a time-related trend.

5. Discussion and conclusions

The natural high variability typical of short environmental records such as peak flow data and the lack of long records has previously hindered the ability of at-site tests to identify clear signals of change in high

river flow across large regions (Mallakpour & Villarini, 2015; Prosdocimi et al., 2014). In this study, we use areal models to pool together the information that directly measure the strength of the evidence a change in peak flows over time across all stations. Using this approach, we find strong evidence for a positive trend in the magnitude of gauged annual maxima of peak river flow in GB. This holds true when different subsets of the available records are analyzed and when using robust regression approaches in the derivation of the test statistic. The signal is clearly detected when all test statistic values across the island are modeled simultaneously in an areal model. These results are in line with those in Brady et al. (2019), in which a similar strength in change in time in near natural catchments was identified using more complex and computationally demanding spatial models. Exploiting the spatial structure of the flow data enhances the trend signal and allows for a clearer inference, thus bridging the previously reported discrepancy between the projected increases in flood risk in GB and the lack of clear signal in the observational peak flow records. Further, the model identifies areas for which the area-specific evidence for a (positive) trend is strong, allowing for a spatial characterization of the potential changes in floods. These areas would be the natural candidates for more in-depth analysis of changes in flood frequencies.

In this study we do not attempt to explain the driving causes that lead to the observed change but rather focus on presenting strong evidence that a change has indeed occurred. The fact that the high flows in the most recent years appear to have on average higher values than those in the past does pose a challenge in terms of whether the full record available at each station should be used when estimating flood frequencies and whether some adjustments should be put in place to account for the fact that estimates obtained using the whole record might underestimate the current flood frequencies (see, e.g., Luke et al., 2017, for a suggestion of such a correction). The approach presented in this study could easily be applied to other parts of the world and other types of environmental data: Pooling the information on the strength of trend at different stations will likely enhance the ability of detecting clearer signals of change across large measuring networks.

References

Alfieri, L., Burek, P., Feyen, L., & Forzieri, G. (2015). Global warming increases the frequency of river floods in Europe. *Hydrology and Earth System Sciences*, *19*(5), 2247–2260. https://doi.org/10.5194/hess-19-2247-2015

Archfield, S. A., Hirsch, R. M., Viglione, A., & Blöschl, G. (2016). Fragmented patterns of flood change across the United States. *Geophysical Research Letters*, *43*, 10,232–10,239. https://doi.org/10.1002/2016GL070590

Barker, L., Hannaford, J., Muchan, K., Turner, S., & Parry, S. (2016). The Winter 2015/2016 floods in the UK: A hydrological appraisal. *Weather*, *71*(12), 324–333. https://doi.org/10.1002/wea.2822

Brady, A., Faraway, J., & Prosdocimi, I. (2019). Attribution of long-term changes in peak river flows in Great Britain. *Hydrological Sciences Journal*, *64*(10), 1159–1170. https://doi.org/10.1080/02626667.2019.1628964

Dankers, R., Arnell, N. W., Clark, D. B., Falloon, P. D., Fekete, B. M., Gosling, S. N., et al. (2014). First look at changes in flood hazard in the inter-sectoral impact model intercomparison project ensemble. *Proceedings of the National Academy of Sciences*, *111*(9), 3257–3261. https://doi.org/10.1073/pnas.1302078110

Do, H. X., Westra, S., & Michael, L. (2017). A global-scale investigation of trends in annual maximum streamflow. *Journal of Hydrology*, *552*, 28–43. https://doi.org/10.1016/j.jhydrol.2017.06.015

Environment Agency (2012). Flood estimation guidelines, Operational Instruction 197_08: Environment Agency.

Fischer, E. M., & Knutti, R. (2014). Detection of spatially aggregated changes in temperature and precipitation extremes. *Geophysical Research Letters*, *41*, 547–554. https://doi.org/10.1002/2013GL058499

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd). United States: Chapman and Hall/CRC.

Gelman, A., & Hill, J. (2012). *Data analysis using regression and multi level/hierarchical models*. New York, NY, USA: Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211.

Giuntoli, I., Villarini, G., Prudhomme, C., Mallakpour, I., & Hannah, D. M. (2015). Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States. *Journal of Geophysical Research: Atmospheres*, *120*, 9138–9159. https://doi.org/10.1002/2015JD023401

Hall, J., Arheimer, B., Borga, M., Brázdil, R., Claps, P., Kiss, A., et al. (2014). Understanding flood regime changes in Europe: A state-of-the-art assessment. *Hydrology and Earth System Sciences*, *18*(7), 2735–2772. https://doi.org/10.5194/hess-18-2735-2014

Hannaford, J. (2015). Climate-driven changes in UK river flows: A review of the evidence. *Progress in Physical Geography: Earth and Environment*, *39*(1), 29–48. https://doi.org/10.1177/0309133314536755

Hannaford, J., Buys, G., Stahl, K., & Tallaksen, L. M. (2013). The influence of decadal-scale variability on trends in long European streamflow records. *Hydrology and Earth System Sciences*, *17*(7), 2717–2733. https://doi.org/10.5194/hess-17-2717-2013

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al. (2013). Global flood risk under climate change. *Nature Climate Change*, *3*(9), 816.

Hodgkins, G. A., Whitfield, P. H., Burn, D. H., Hannaford, J., Renard, B., Stahl, K., et al. (2017). Climate-driven variability in the occurrence of major floods across North America and Europe. *Journal of Hydrology*, *552*, 704–717. https://doi.org/10.1016/j.jhydrol.2017.07.027

IPCC (2012). Managing the risks of extreme events and disasters to advance climate change adaptation. A special report of Working Groups I and II of the Intergovernmental Panel on Climate Change: Intergovernmental panel on climate change (pp. 594).

Kay, A. L., Crooks, S. M., Davies, H. N., & Reynard, N. S. (2014). Probabilistic impacts of climate change on flood frequency using response surfaces I: England and Wales. *Regional Environmental Change*, *14*(3), 1215–1227. https://doi.org/10.1007/s10113-013-0563-y

Kay, A. L., Crooks, S. M., Davies, H. N., & Reynard, N. S. (2014). Probabilistic impacts of climate change on flood frequency using response surfaces II: Scotland. *Regional Environmental Change*, *14*(3), 1243–1255. https://doi.org/10.1007/s10113-013-0564-x

Kendon, M., Marsh, T., & Parry, S. (2013). The 2010-2012 drought in England and Wales. *Weather*, *68*(4), 88–95. https://doi.org/10.1002/wea.2101

Lamb, R., Faulkner, D., & Zaidman, M. D. (2009). Fluvial design guide—Chapter 2: Hydrology: Environment agency. http://evidence.environment-agency.gov.uk/FCERM/Libraries/Fluvial_Documents/Fluvial_Design_Guide_-_Chapter_2.sflb.ashx

Lawson, A. B. (2013). *Bayesian disease mapping: Hierarchical modeling in spatial epidemiology*: Chapman and Hall/CRC.

Luke, A., Vrugt, J. A., AghaKouchak, A., Matthew, R., & Sanders, B. F. (2017). Predicting nonstationary flood frequencies: Evidence supports an updated stationarity thesis in the United States. *Water Resources Research*, *53*, 5469–5494. https://doi.org/10.1002/2016WR019676

Mallakpour, I., & Villarini, G. (2015). The changing nature of flooding across the central United States. *Nature Climate Change*, *5*(3), 250.

Marsh, T., & Lees, M. (1985). The 1984 drought. Wallingford, UK: Natural Environment Research Council, Institute of Hydrology.

Mediero, L., Kjeldsen, T. R., Macdonald, N., Kohnova, S., Merz, B., Vorogushyn, S., et al. (2015). Identification of coherent flood regions across Europe by using the longest streamflow records. *Journal of Hydrology*, *528*, 341–360. https://doi.org/10.1016/j.jhydrol.2015.06.016

Murphy, J. M., Sexton, D. M. H., Jenkins, G. J., Booth, B. B. B., Brown, C. C., Clark, R. T., et al. (2009). UK climate projections science report: Climate change projections. Exeter: Met Office Hadley Centre.

National River Flow Archive (2014). Hydrometric areas for Great Britain and Northern Ireland: NERC Environmental Information Data Centre. https://doi.org/10.5285/1957166d-7523-44f4-b279-aa5314163237

National River Flow Archive (2018). Walling-ford, UK: National River Flow Archive: NERC/Centre for Ecology & Hydrology. https://nrfa.ceh.ac.uk

Prosdocimi, I., Kjeldsen, T. R., & Svensson, C. (2014). Non-stationarity in annual and seasonal series of peak flow and precipitation in the UK. *Natural Hazards and Earth System Sciences*, *14*(5), 1125–1144. https://doi.org/10.5194/nhess-14-1125-2014

Renard, B., Garreta, V., & Lang, M. (2006). An application of Bayesian analysis and Markov chain Monte Carlo methods to the estimation of a regional trend in annual maxima. *Water Resources Research*, *42*, W12422. https://doi.org/10.1029/2005WR004591

Renard, B., Lang, M., Bois, P., Dupeyrat, A., Mestre, O., Niel, H., et al. (2008). Regional methods for trend detection: Assessing field significance and regional consistency. *Water Resources Research*, *44*, W08419. https://doi.org/10.1029/2007WR006268

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Sayers, P. B., Horritt, M, Penning-Rowsell, E, & McKenzie, A. (2015). Climate Change Risk Assessment 2017: Projections of future flood risk in the UK (pp. 126). London: Committee on Climate Change.

Slater, L. J., & Villarini, G. (2016). Recent trends in U.S. flood risk. *Geophysical Research Letters*, *43*, 12,428–12,436. https://doi.org/10.1002/2016GL071199

Svensson, C., Hannaford, J., Kundzewicz, Z. W., & Marsh, T. J. (2006). Trends in river floods: Why is there no clear signal in observations? *IAHS Publications-Series of Proceedings and Reports*, *305*, 1–18.

United Nations Office for Disaster Risk Reduction (UNISDR) (2015). Making development sustainable: The future of disaster risk management. Global Assessment Report on Disaster Risk Reduction: United Nations Office for Disaster Risk Reduction (pp. 316).

Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. New York: Springer Science & Business Media.

Vogel, R. M., Rosner, A., & Kirshen, P. H. (2013). Brief communication: Likelihood of societal preparedness for global change: trend detection. *Natural Hazards and Earth System Sciences*, *13*(7), 1773–1778. https://doi.org/10.5194/nhess-13-1773-2013

Vogel, R. M., Yaindl, C., & Walter, M. (2011). Nonstationarity: Flood magnification and recurrence reduction factors in the United States. *JAWRA Journal of the American Water Resources Association*, *47*(3), 464–474. https://doi.org/10.1111/j.1752-1688.2011.00541.x

RESEARCH LETTER
10.1029/2019GL085142

Key Points:

- We propose a novel approach to regional detection of trends in measured series based on areal models
- We detect a clear signal that peak flows magnitudes are increasing over time in Great Britain
- These changes are still found when different periods of record are analyzed, with an accelerated upward trend from 1980 onward

Supporting Information:

- Supporting Information S1
- Data Set S1
- Data Set S2

Correspondence to:
I. Prosdocimi,
Ilaria.Prosdocimi@unive.it

Citation:
Prosdocimi, I., Dupont, E., Augustin, N., Kjeldsen, T. R., Simpson, D., & Smith, T. (2019). Areal models for spatially coherent trend detection: The case of British peak river flows. *Geophysical Research Letters*, *46*, 13,054–13,061. https://doi.org/10.1029/2019GL085142

Received 25 AUG 2019
Accepted 28 OCT 2019
Accepted article online 9 NOV 2019
Published online 29 NOV 2019

Acknowledgments
Emiko Dupont is supported by a scholarship from the EPSRC Centre for Doctoral Training in Statistical Applied Mathematics at Bath (SAMBa), under the project EP/L015684/1. Ilaria Prosdocimi was supported by an NPIF Innovation Fellowship from the Natural Environment Research Council (NERC), under the project NE/R013152/1. The WINFAP data files and the hydrometric area shapefiles on which the analysis is based can be retrieved at the U.K. National River Flow Archive (NRFA, https://nrfa.ceh.ac.uk/). The HadUK-Grid can be retrieved at the Met Office (https://www.metoffice.gov.uk/research/climate/maps-and-data/). The authors thank the NRFA and the measuring authorities for making the river flow data available. The authors also thank the Met Office for making the data for the climate average of the UK available. The R scripts used to read, select, and analyze the data as well as creating all figures in the manuscript are provided as supporting information and provided online (at http://doi.org/10.5281/zenodo.3497404).

Supporting Information for "Areal models for spatially coherent trend detection: the case of British peak river flows"

Ilaria Prosdocimi[1,2], Emiko Dupont[2], Nicole H. Augustin[2], Thomas R. Kjeldsen[3], Dan P. Simpson[4], Theresa R. Smith[2]

[1]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy

[21]Department of Mathematical Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, UK

[3]Department of Architecture and Civil Engineering, University of Bath, Claverton Down, Bath, BA2 7AY, UK

[4]Department of Statistical Sciences, University of Toronto, Sidney Smith Hall, 100 St. George St.,Toronto, Ontario M5S 3G3, Canada

Contents of this file

1. Text S1 to S6
2. Figures S1 to S8
3. Tables S1 to S2

Additional Supporting Information (Files uploaded separately)

1. `dataPreparation.R`: R file to create the datasets underpinning the analysis

---

October 22, 2019, 3:53pm

2. `dataAnalyisAndFigures.R`: R file to carry out the analysis and create the Figures and Tables presented in the paper and in the Supplementary Information

The two are files can also be found at http://doi.org/10.5281/zenodo.3497404

Introduction

The supplementary information is sub-divided in 6 main topics, each within a Text block which introduces the relevant Figures and Tables. Text S1 presents some additional information on the spatio-temporal coverage of the river flow data used in the study. Text S2 gives some additional background on the derivation of the test statistics used in the analysis. Text S3 gives some additional information on the climatology of Great Britain and on the similarity in climate within hydrometric areas. Text S4 shows the estimated posterior mean and credible intervals for the areal models derived in each sub-analysis on the long common period. Text S5 shows summary statistics for the estimated posterior of parameters of the areal models, including the variance components. Text S6 presents the results obtained when using the test statistics derived using robust regression.

Text S1.

Available flow records

Figure S1 and S2 show the temporal evolution of the record availability across Great Britain. The drop in data availability after the early 2000s visible in Figure S1 is due to a delay in the processing of the gauged data by the Scottish measuring authorities, as evident in Supplementary Figure S2. Figure S2 also shows what portions of Great Britain and of the hydrometric areas are located in England, Wales or Scotland.

Text S2.

Test statistic for a linear trend

We used a simple linear regression model applied to the logarithm of peak flow with time as an explanatory variable to assess whether or not there are trends in the magnitude of peak flow at each station:

$$\log(Q_{wy}) = \alpha + \beta\ wy + \varepsilon_{wy} \tag{1}$$

where $\varepsilon_{wy} \sim N(0, \sigma^2)$ for every wy, and α, β and σ are parameters which need to be estimated. In the notation above wy represents the water year in which the high flow value Q_{wy} was measured, and it varies between $(wy_1, \ldots, wy_n)$, with wy_1 being the first year in the record and wy_n being the last year in the record. The Q_{wy} variable represents the annual maximum peak flow value (measured in m^3/s). This model was introduced by Vogel, Yaindl, and Walter (2011) and applied to the British data in Prosdocimi, Kjeldsen, and Svensson (2014), where it was found to fit well the British data.

At each station, the trend parameter β is estimated as

$$\hat{\beta} = \rho(Q, wy)\frac{sd(Q)}{sd(wy)}$$

where $\rho(Q, wy)$ is the sample correlation coefficient between peak river flow series and the time variable, while $sd(Q)$ and $sd(wy)$ are, respectively, the sample standard deviation of the peak river flow series and the time variable. The standard deviation of the trend estimate can be derived using standard results for linear regression models as:

$$sd(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(wy_i - \overline{wy})^2}}$$

where $\hat{\sigma}$ indicates the estimate for regression error, which can be derived as $\hat{\sigma} = sd(Q) * \sqrt{(1 - \rho(Q, wy)^2)}$, and $\overline{wy}$ indicates the average value of the time variable. A test statistic for trend, i.e. for the system of hypothesis

$$H_0 : \beta = 0 \qquad VS \qquad H_1 : \beta \neq 0$$

can therefore be constructed as

$$T = \frac{\hat{\beta}}{sd(\hat{\beta})}.$$

According to the statistical theory, under the null hypothesis the test statistic follows a T-distribution with $(n - 2)$ degrees of freedom, which resembles closely the standard normal distribution for relatively large values of n. By calculating the necessary quantities at each station separately a set of test statistic values $(T_1, \ldots, T_{640})$ is derived. Each test statistic is a summary of the strength of the time trend at each station. The set of all test statistics is then modelled simultaneously by means of the areal model presented in the main text. A second set of test statistics $(T_1^R, \ldots, T_{640}^R)$ is also derived for the same model in equation (1) and the same system of hypothesis for trend using a robust approach to linear regression estimation as presented in Yohai (1987). Under the null hypothesis of a null slope these test statistics are also asymptotically normally distributed. Using a robust approach ensures that the estimate of the slope in the linear model is not unduly affected by large flow events in the series. Although the main analysis focuses on the test statistics derived from standard liner models, the set of robust test statistics is used to ensure that the reported findings are not unduly influenced by some of the larger events recorded in the latter years in the records.

According to the statistical theory underlying the construction of the test, if the null hypothesis of no trend was true across all stations and the tests at each station were independent, we would expect the values of the test statistic across all stations to behave approximately like a standard normal distribution. Examining the overall distribution of test statistic values for the complete dataset in Figure S3, a misalignment to the theory is evident: the overall mean and standard deviation are found to be 0.733 and 1.36 and the histogram shown in the left panel of Figure S3 is clearly different from what we would expect to see if the test values behaved according to the standard normal distribution. This is also true for the case in which the 298 records with the long common time period (1976-2016) are used to calculate the test statistics (right panel). Possibly, this misalignment is a consequence of the fact that hypothesis of no trend does not hold overall and that the tests performed at each station are not independent, since nearby stations might gauge the same river and are affected by the same climate and weather.

Text S3.

Climatology of Great Britain and Hydrometric Areas

In the areal model introduced in the main text it is assumed that the evidence for trend is similar within each hydrometric area (and not that the flow records within each area are homogeneous). One of the reasons which justifies assuming that stations within each hydrometric areas can be expected to be experiencing a similar trend, is that they are located in similar climate, would experience similar weather and would in some cases be gauging different sub-catchments of the same river. To investigate how similar stations

within hydrometric areas might be from a climatological point of view, a map of the long term temperature and rainfall climatology across Great Britain is displayed in Figure S4, together with two measures of seasonality of the annual maximum flow at each station. These two measures are, the median day of peak flow and the modal (i.e. most frequent) month of peak flow at each station. These latter measures were derived from the annual maxima records: the water year in the UK begins on October 1st and all measures are derived accordingly, so a median day of flow equal to 1 would indicate that the median day of peak flow at the station is October 1st. The long term climatologies, derived from measurements in the years between 1981 and 2010, are provided at a 1 km grid resolution by the Met Office (2018). The presence of mountainous ranges is clearly detectable in both the rainfall and temperature maps. These ranges divide the country in a cooler and wetter north-western part, characterised also by a more marked winter seasonality and a drier, warmer part in which later peak flows are more common. Although for the HAs which contain the higher mountain peaks there might be some differences in the climatologies, with heavier rainfalls and cooler temperatures on the mountain tops, these HAs tend to include stations which are directly connected hydrologically, i.e. that gauge subcatchments of the same river. Conversely, the HAs in which stations might be effectively gauging different river courses tend to be located in the southern part of the country, where the climate is more homogeneous.

Text S4.

Time evolution of evidence for trend across the long common period of record

Figure S5 and S6 show the evolution of the estimated posterior mean for each area specific test statistic value when using different 31-year long subsets of the data and the 41-year long subset in the period 1976-2016 (the long common period of record).

Text S5.

Inference for variance components in the areal models

In Table S1 key summary statistics for the marginal posterior distribution of the overall trend value and the variance components of the areal model presented in equation (1) of the main text are shown. The fairly high estimated values of the σ_H^2 variance component highlight the need for the area-specific effects h_j to be included in the model since their inclusion explains a large part of the variability in the data. This is also true for the model based on the records with a long common time period (1976-2016), for which key summary statistics are shown in Table S2.

Text S6.

Inference when the test statistics are derived using robust regression approaches

Figure S7 is structured as Figure 1 in the main text although the test statistics displayed on the left hand side map and on which the areal model is fitted corresponds to the test statistics derived using robust regression approaches rather than standard linear regression. The general findings using the alternative robust regression approaches are fairly similar to those found when using results based on the standard linear models. Some

variations are visible in the test statistic values and the properties of the posterior distribution in each HA. In particular for less areas is the 90% credible interval found to only include positive values, and in one case, HA 39, the Thames, the 90% credible interval is found to only include negative values. The 90% credible interval for the overall trend effect μ is (0.47, 0.73) which is slightly wider than that reported in the main text, but still does clearly not include the null value. The indication of strong evidence in favour of an increasing trend in river flow is still present.

This is still true also when the analysis is carried out on the long period of record and its subsets. Figure S8 is structured in the same way as Figure 3 in the main text. Although there is less variability in the posterior distributions in the different sub-periods, it is still clear that the overall trend is generally positive with no 90% credible interval in any sub-period containing negative values.

References

Met Office. (2018). *Haduk-grid gridded and regional average climate observations for the uk. centre for environmental data analysis* (Tech. Rep.). Centre for Environmental Data Analysis. Retrieved from `http://catalogue.ceda.ac.uk/uuid/4dc8450d889a491ebb20e724debe2dfb`

Prosdocimi, I., Kjeldsen, T., & Svensson, C. (2014). Non-stationarity in annual and seasonal series of peak flow and precipitation in the UK. *Natural Hazards and Earth System Sciences*, *14*(5), 1125–1144. doi: 10.5194/nhess-14-1125-2014

Vogel, R. M., Yaindl, C., & Walter, M. (2011). Nonstationarity: Flood Magnification and Recurrence Reduction Factors in the United States. *JAWRA Journal of the*

American Water Resources Association, *47*(3), 464–474. doi: 10.1111/j.1752-1688.2011.00541.x

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, *15*(2), 642–656.

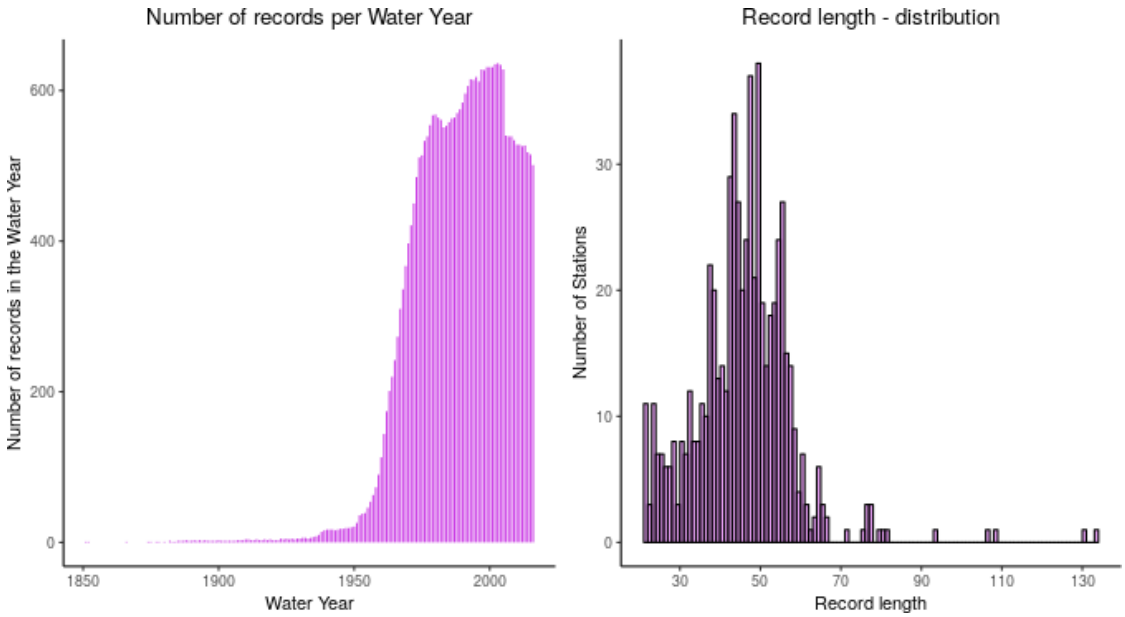


Figure S1. Left panel: number of records available per year. Right panel: histogram of the record length available at each station.

October 22, 2019, 3:53pm

:

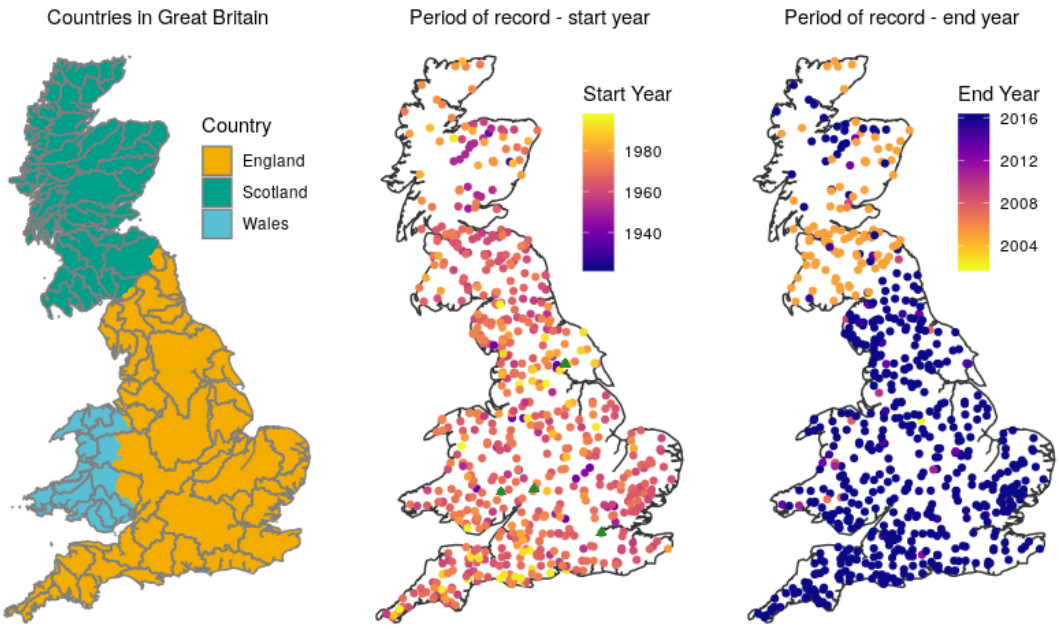


Figure S2. Left panel: division of Great Britain according to the country, including hydrometric areas (HA). Central panel: first year of valid flow measurements at each station (green triangles indicate stations which began recording before 1916). Right panel: last year of valid flow measurements at each station.

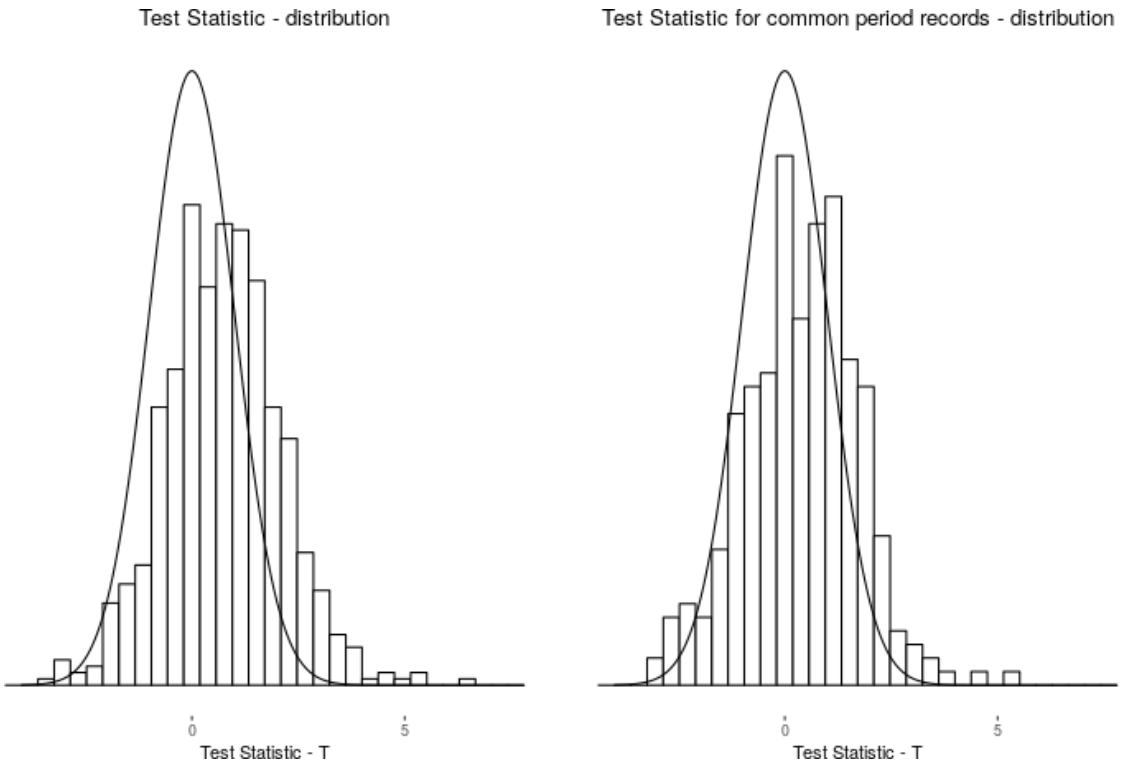


Figure S3. Histogram of the test statistics for all stations and the pdf of the standard normal distribution. Left panel: full dataset; right panel: long common time period (1976-2016) dataset

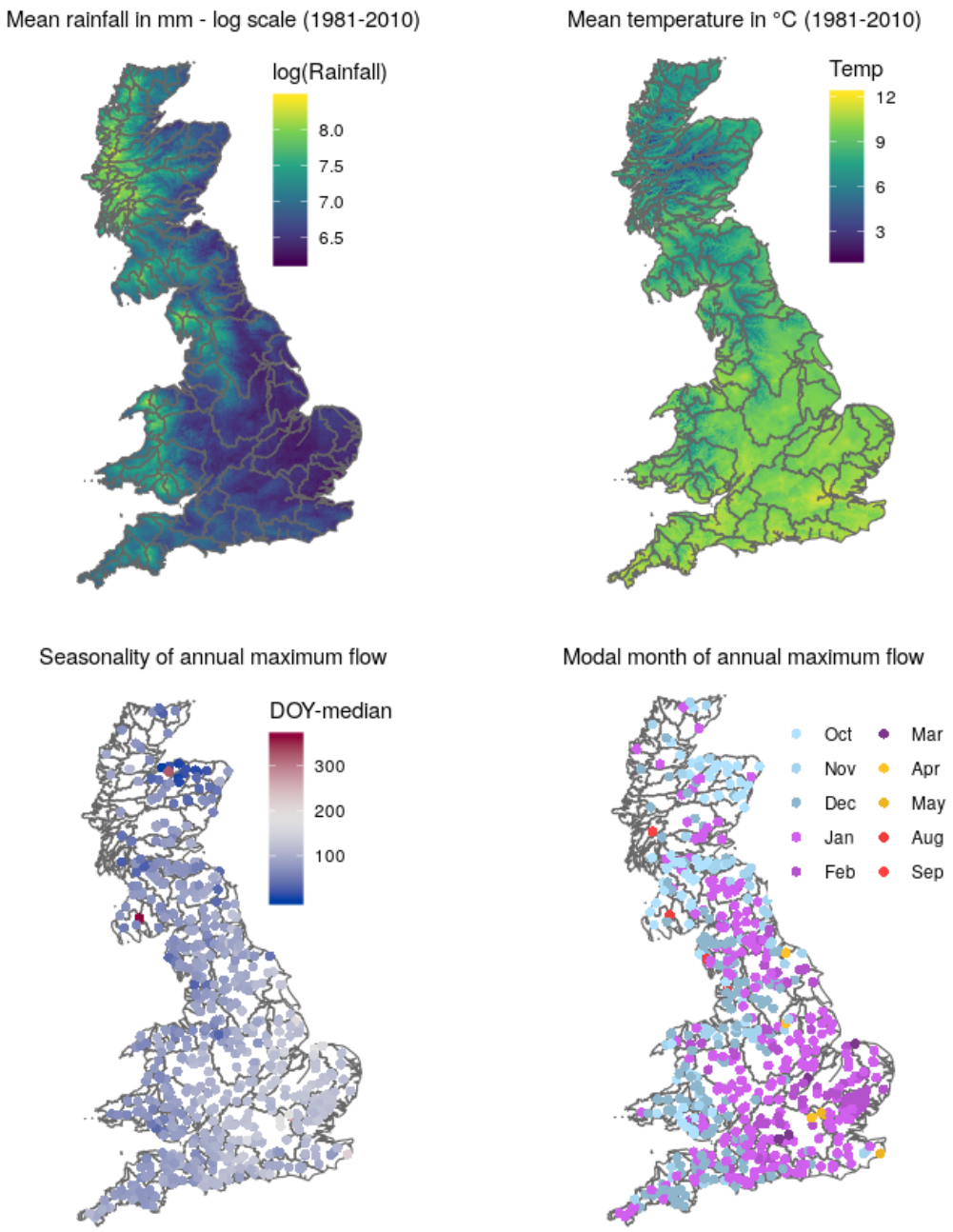


Figure S4. Top left panel: rainfall climatology (1981-2010) - mm, log scale. Top right panel: temperature climatology (1981-2010) - °C. Bottom left panel: seasonality of annual maxima flow, median day of peak flow. Bottom right panel: seasonality of annual maxima flow, modal month of peak flow.

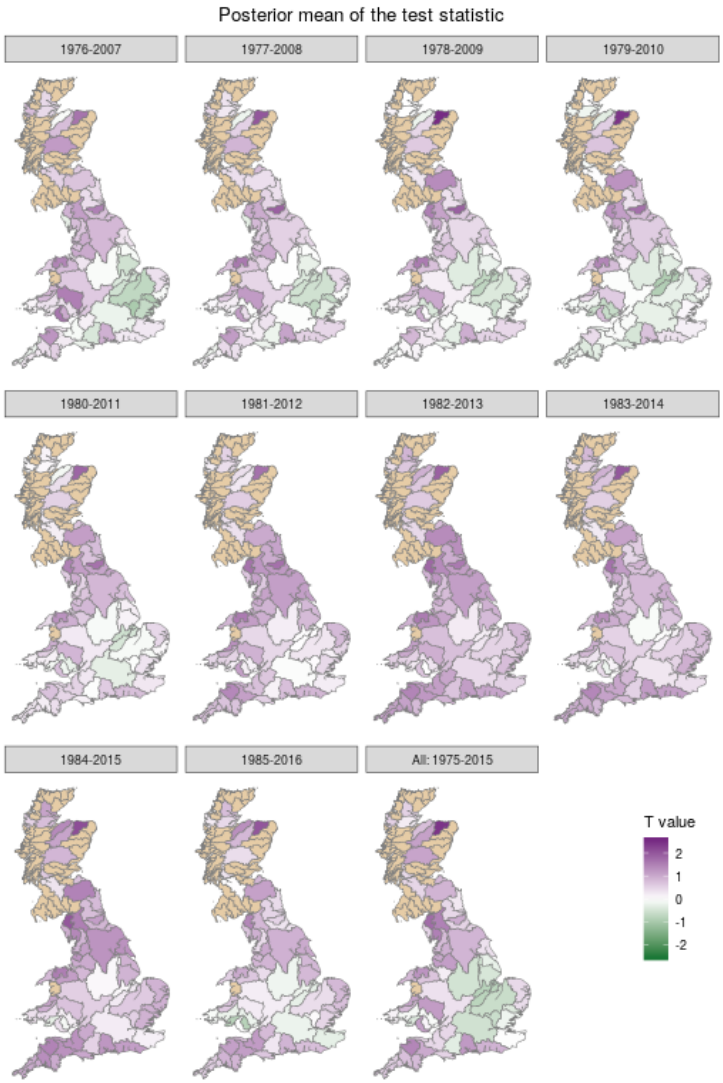


Figure S5. Estimated posterior mean for each area specific test statistic value when using different 31-year long subsets of the data and the 41-year long subset in the period 1976-2016.

:

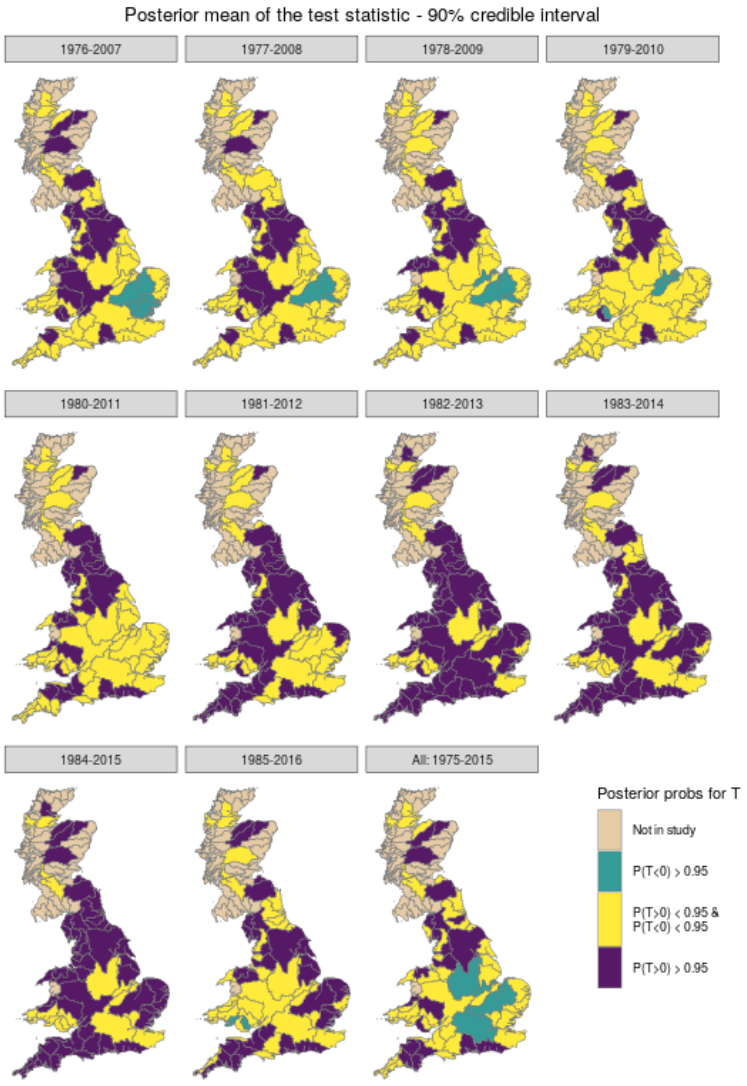


Figure S6. Summarised information for the 90% credible interval of each area specific test statistic value when using different 31-year long subsets of the data and the 41-year long subset in the period 1976-2016.

October 22, 2019, 3:53pm

Table S1. Inference for key parameters in the areal model

	0.025 perc.	0.05 perc.	median	0.95 perc.	0.975 perc.	Mean
μ	0.6144	0.6410	0.7772	0.9142	0.9410	0.7777
σ_T^2	1.3665	1.3911	1.5311	1.6886	1.7200	1.5345
σ_H^2	0.1723	0.1881	0.2979	0.4610	0.4996	0.3077

Table S2. Inference for key parameters in the areal model when using the subset of records with long common time period (1976-2016)

	0.025 perc.	0.05 perc.	median	0.95 perc.	0.975 perc.	Mean
μ	0.2675	0.3076	0.5122	0.7186	0.7595	0.5130
σ_T^2	0.9149	0.9403	1.0894	1.2678	1.3046	1.0950
σ_H^2	0.3812	0.4137	0.6391	0.9764	1.0562	0.6597

:

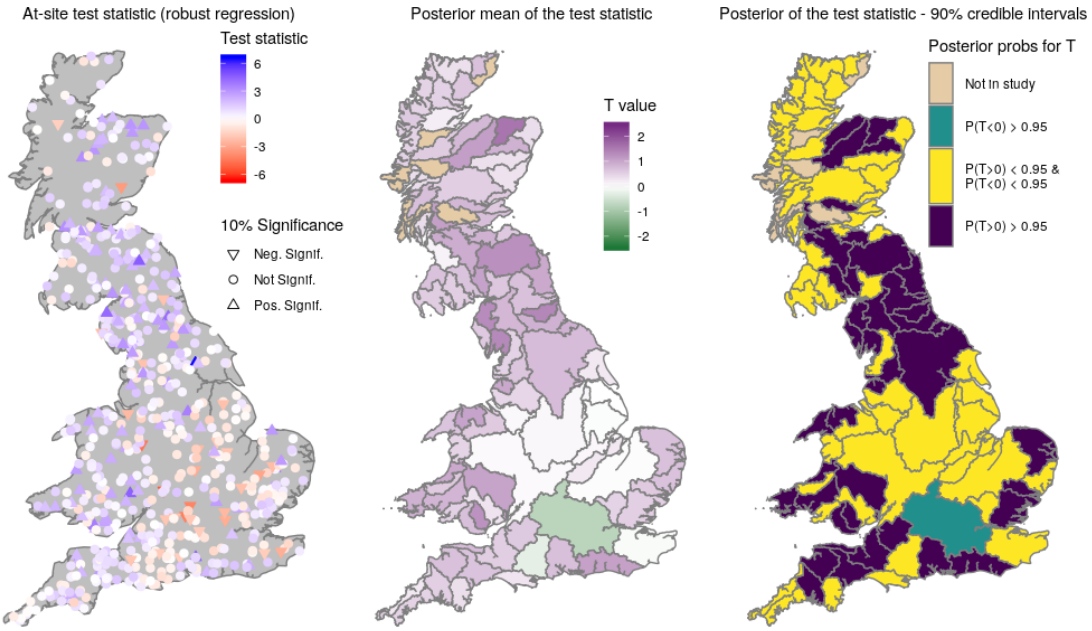


Figure S7. Left panel: at-site test statistic derived using robust regression and significance at 10% significance level for all stations. Central panel: estimated posterior mean derived from the proposed areal model for each area specific test statistic value. Right panel: summarised information for the 90% credible interval for each area specific test statistic value. Test statistics included in the areal model derived using robust regression.

October 22, 2019, 3:53pm

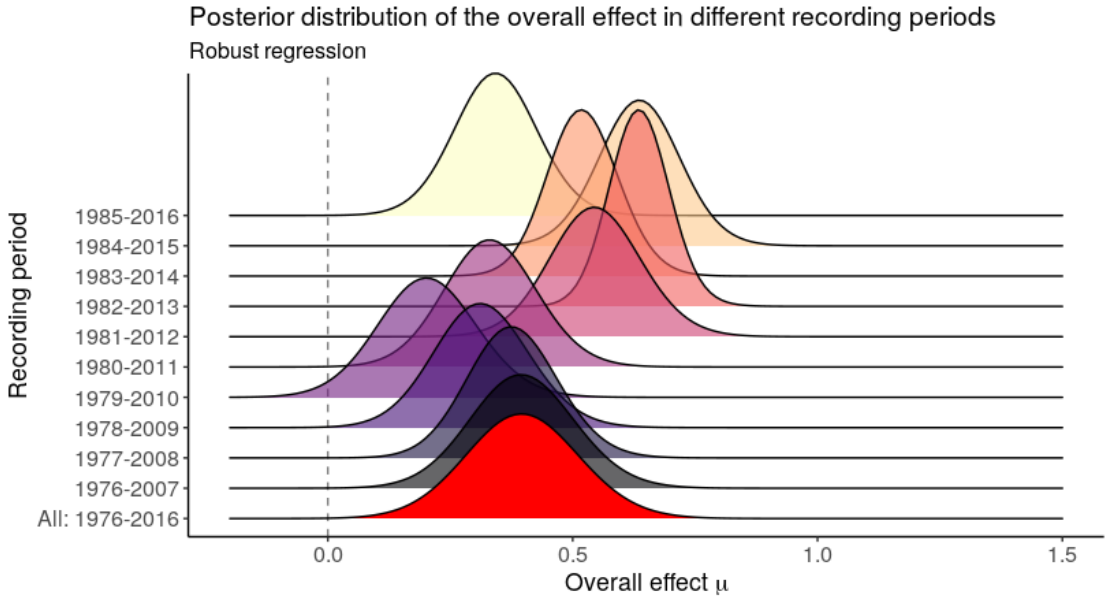


Figure S8. Estimated posterior distributions of μ when using different 31-year long subsets and the 41-year long subset in the period 1976-2016. Test statistics included in the areal model derived using robust regression.

Closing remarks for Paper 2

Partial pooling is a well-known method for sharing information between different spatial locations. However, modelling the test statistic for the enhancement of a trend signal in this way is novel and provides a relatively simple method for assessing the overall time trend across the geographical study region as well as the regional variation in this trend signal. The straightforward implementation means that it could be used, for example, as a monitoring tool for observing changes in flood risk as well as identifying the geographical regions that are of most concern.

Using the test statistic, rather than directly modelling the flow records, is advantageous here as it captures exactly the information of interest, namely, the strength of the evidence for a trend as well as the trend direction. Moreover, the test statistic has (approximately) the same standard normal distribution at all locations and, therefore, test statistics from different stations are directly comparable. In contrast, flow records vary significantly in scale from station to station (as the rivers they measure can vary greatly in size) and such records would therefore need to be standardised in some way to be able to model them together in the same spatial model.

Trend detection across large measuring networks is of interest in many areas of environmental statistics and our method could easily be transferred to other applications. We note that, although the spatial model used in the paper was a simple iid random effects model, the model could also be adjusted to take account of more complex spatial variation. For example, we also fitted the model with an additional ICAR (Intrinsic Conditional Autoregressive) random effect to reflect spatial correlation between HAs, however, there was no clear evidence in the data for this additional structure. The geographical partition used for the spatial model is also a choice that will depend on the structure of the data and the intended use for the model output, in particular, the spatial scales of interest. Here we used HAs as they are a well-established way of partitioning Great Britain for hydrometric analysis, and the resulting model allows sufficient data pooling to be able to detect a clear signal while still capturing the regional variation. However, other partitions could also have been used.

Chapter 5

Conclusions

This chapter summarises our overall conclusions from the thesis and sets out some ideas for future research.

Overall conclusions

Spatial models allow data collected at different spatial locations to be modelled together. These models use spatial random effects to reflect the residual spatial correlation structure in the data, resulting in fitted values that are in some way smoothed across the spatial domain. Spatial models are becoming an increasingly common tool in many areas of applied statistics. Paper 2 in this thesis illustrates how spatial modelling can be used for data pooling. When flow data for rivers in Great Britain are modelled separately at each gauging station, the short data records make it difficult to recognise a clear trend. By modelling the test statistic for a trend in a spatial model, information from different spatial locations can be shared to enhance the statistical signal. Thus, we are able to detect, for the first time, a significant positive trend in flood risk over time and, moreover, identify the geographical areas that have the strongest trend. Using partial pooling in this way to detect spatially coherent trend signals is novel and the method could easily be transferred to other applications.

The growing use of spatial models has generated more interest in understanding some of the consequences that spatial random effects have on statistical inference, in particular, the problem of unreliable covariate effect estimates due to spatial confounding. One of the main contributions of the work in this thesis is to provide a relatively accessible theoretical explanation for why this problem arises. In practice, spatial confounding is usually detected in applications when a null model (with no spatial effects) and a spatial model are fitted to the same data but give noticeably different results for the covariate effect estimates. Using the thin plate spline formulation of the spatial model we explicitly analyse the expressions for the effect estimates in these models using simple linear algebra. The perhaps surprising conclusion from our investigation is that neither of these models can be relied upon to estimate the correct effects, although for different reasons.

While some authors in the literature have noted that the estimates in the null model are biased when unmeasured spatial effects are present, there seems to have been a general misconception that the null effects are "correct" and should be protected from interference from any spatial terms added to the model. This has lead to the widespread use of methods such as RSR which include spatial effects but with added constraints designed to eliminate or reduce collinearity with the covariates, thereby recovering the estimate in the null model. However, this approach has recently come under considerable criticism (see, for example, Khan and Calder [2020]) and, as we have also shown, RSR creates rather than removes bias in covariate effect estimates in the presence of unmeasured spatial effects.

For the spatial model, we see that spatial confounding bias depends on the structure of the covariate of interest. A common assumption in the spatial confounding literature is that the covariate has a spatial correlation structure, i.e. it is fully determined by spatial location. In this case, there is essentially no information to distinguish the covariate from the spatial effects and this unidentifiability leads to bias in its effect estimate. However, our analysis

focusses on a situation which is often the case in practice, namely, where the covariate has some non-spatial information as well. We show that, in this case, the bias is not caused by unidentifiability, but arises because the smoothing applied to the spatial part of the model can have a disproportionate effect on the covariate part. In Paper 1 we propose a novel easily implementable method, spatial+, for dealing with this bias. Spatial+ is defined by replacing the covariates in the spatial model by their residuals after spatial dependence is regressed away. In this paper, we use asymptotic analysis of the resulting effect estimates as well as simulations to show that the method works. Our analysis in Chapter 2 also provides some intuition for the method. We see that the spatial part of the covariate is actually unnecessary for identifying the covariate effect in the spatial model, and moreover, by removing this part we obtain an effect estimate that is decoupled from the spatial effect and therefore much less sensitive to smoothing. This means that the covariate effect estimate stays broadly unbiased.

Future work

The approach of Rice, Chen and Shiau that we have generalised to higher dimensions provides theoretical backing for the spatial+ method in the particular setting where the spatial model is formulated using a thin plate spline. However, the intuition behind the method, namely, the decoupling of the covariate effect estimate from the spatial effect in the spatial model obtained by removing spatial dependence from the covariate part of the model matrix, is an idea that can be transferred to spatial models more generally. The approach of spatial+ is novel as existing methods for dealing with spatial confounding typically adapt or constrain the spatial part of the model rather than the covariate part which, as we have seen, tends to create rather than remove bias. In some settings, the spatial+ method may be directly transferable. For example, the commonly used Gaussian Markov random fields (GMRFs) modelled by the stochastic partial differential equation (SPDE) approach [Lindgren et al., 2011] can be understood in the language of smoothing penalties in a similar way to thin plate splines [Miller et al., 2020]. While spatial+ is therefore likely to work, the smoothing penalties applied in other settings are not identical to those of thin plate splines and the method would need to be tested and verified with any new implementations.

In our analysis we assumed that the spatial pattern of the covariates in the spatial model could be identified using the Gaussian model (2.11). However, in some practical applications, covariates in the linear predictor could have non-Gaussian distributions and, in that case, it is unclear how to remove the spatial dependence of the covariates in the model matrix in order to apply spatial+. Another assumption of the approach is that the spatial model is identifiable. However, there may be situations where, despite the assumption of non-spatial information in the covariate, the spatial model is close to being unidentifiable. For example, in a discrete space model with an ICAR random effect modelling spatial dependence between n separate regions, if there is only one observation per region, then the rank of the spatial basis is the same as the sample size n. In that case, it may be harder to reliably identify spatial residuals and the spatial+ method may not be directly applicable. It would be interesting to explore these issues to see if there are still elements of the spatial+ methodology that can be adapted and applied in cases where the current assumptions for the method do not hold.

Although most literature on spatial models and, in particular, spatial confounding consider only linear covariate effects, we have seen that confounding problems are also clearly present for non-linear covariate effects. An advantage of the thin plate spline formulation is that it is easy to implement both linear and non-linear covariate effects in

this setting using the GAM framework. The approach of spatial+, however, does not extend directly to this case as the method relies on linearity of the covariate effects for the effect of the spatial residuals to coincide with the original covariate effect. In some cases, it may be possible to transform a covariate variable in such a way that the effect is broadly linear, making it possible to use spatial+. But otherwise, something more sophisticated would need to be developed.

Finally, our analysis has focussed on spatially induced bias in covariate effect estimates, however, another aspect of spatial confounding is the effect of spatial collinearity on the variance of the estimates. Reich et al. [2006] suggests that collinearity in the spatial model causes variance inflation compared to the null model, however, this is in fact not always the case as their analysis assumes a relationship between the estimated scale parameters in the two models that does not generally hold. Results for RSR [Khan and Calder, 2020, Hanks et al., 2015] show that the credible intervals obtained using this method tend to be inappropriately small, leading to elevated levels of Type-S errors (the Bayesian analogue of Type-1 errors). A natural next step for our analysis of the spatial+ approach would be to investigate, both theoretically and through simulations, how the variance of estimates behaves in this model. With a better understanding of this, we may also be able to develop some diagnostics or tests that could help practioners assess when spatial confounding problems are present and should be dealt with in a given application.

Bibliography

A. Adin, T. Goicoa, J. Hodges, P. Schnell, and M. Ugarte. Alleviating confounding in spatio-temporal areal models: different proposals, different results. *arXiv preprint arXiv:2003.01946*, 2020.

H. Chen and J.-J. H. Shiau. A two-stage spline smoothing method for partially linear models. *Journal of Statistical Planning and Inference*, 27(2):187–201, 1991.

D. G. Clayton, L. Bernardinelli, and C. Montomoli. Spatial correlation in ecological analysis. *International Journal of Epidemiology*, 22(6):1193–1202, 1993.

A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

E. M. Hanks, E. M. Schliep, M. B. Hooten, and J. A. Hoeting. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254, 2015.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

J. S. Hodges and B. J. Reich. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334, 2010.

J. Hughes and M. Haran. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159, 2013.

K. Khan and C. A. Calder. Restricted spatial regression methods: Implications for inference. *Journal of the American Statistical Association*, pages 1–13, 2020.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

D. L. Miller, R. Glennie, and A. E. Seaton. Understanding the stochastic partial differential equation approach to smoothing. *Journal of Agricultural, Biological and Environmental Statistics*, 25(1): 1–16, 2020.

W. S. Nobre, A. M. Schmidt, and J. B. Pereira. On the effects of spatial confounding in hierarchical models. *International Statistical Review*, 2020.

C. J. Paciorek. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 107, 2010.

G. L. Page, Y. Liu, Z. He, and D. Sun. Estimation and prediction in the presence of spatial confounding for spatial linear models. *Scandinavian Journal of Statistics*, 2017.

J. B. Pereira, W. S. Nobre, I. F. Silva, A. M. Schmidt, et al. Spatial confounding in hurdle multilevel beta models: the case of the brazilian mathematical olympics for public schools. *Journal of the Royal Statistical Society Series A*, 183(3):1051–1073, 2020.

I. Prosdocimi, T. Kjeldsen, and C. Svensson. Non-stationarity in annual and seasonal series of peak flow and precipitation in the uk. *Natural Hazards and Earth System Sciences*, 14:1125–1144, 2014.

I. Prosdocimi, E. Dupont, N. H. Augustin, T. R. Kjeldsen, D. P. Simpson, and T. R. Smith. Areal models for spatially coherent trend detection: the case of british peak river flows. *Geophysical Research Letters*, 46(22):13054–13061, 2019.

B. J. Reich, J. S. Hodges, and V. Zadnik. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206, 2006.

B. J. Reich, S. Yang, Y. Guan, A. B. Giffin, M. J. Miller, and A. G. Rappold. A review of spatial causal inference methods for environmental and epidemiological applications. *arXiv preprint arXiv:2007.02714*, 2020.

J. Rice. Convergence rates for partially splined models. *Statistics & probability letters*, 4(4):203–208, 1986.

S. H. Sørbye, J. B. Illian, D. P. Simpson, D. Burslem, and H. Rue. Careful prior specification avoids incautious inference for log-gaussian cox point processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):543–564, 2019.

H. Thaden and T. Kneib. Structural equation models for dealing with spatial confounding. *The American Statistician*, 72(3):239–252, 2018.

F. Utreras. Natural spline functions, their associated eigenvalue problem. *Numerische Mathematik*, 42(1):107–117, 1983.

F. I. Utreras. Convergence rates for multivariate smoothing spline functions. *Journal of approximation theory*, 52(1):1–27, 1988.

G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

S. N. Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.